Technical Report 1296

Tier One Performance Screen Initial Operational Test and Evaluation: 2010 Annual Report

Deirdre J. Knapp (Ed.)
Human Resources Research Organization

Tonia S. Heffner (Ed.)
U.S. Army Research Institute

October 2011

United States Army Research Institute for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS, Ph.D.
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Kate LaPort, U.S. Army Research Institute
Irwin Justin Jose, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
October 2011	Interim	August 2009 to December 2010

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Tier One Performance Screen Initial Operational Test and Evaluation: 2010 Annual Report	W91WAS-09-D-0013
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S)	5c. PROJECT NUMBER A790
Deirdre J. Knapp (Ed.) (Human Resources Research Organization) and Tonia S. Heffner (Ed.) (U.S. Army Research Institute)	5d. TASK NUMBER 329
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Human Resources Research Organization 66 Canal Center Plaza, Suite 700 Alexandria, Virginia 22314	FR11-40
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: DAPE-ARI-RS 2511 Jefferson Davis Highway Arlington, VA 22202-3926	ARI
	11. MONITOR REPORT NUMBER Technical Report 1296

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contracting Officer's Representative and Subject Matter Expert POC: Dr. Tonia Heffner

14. ABSTRACT *(Maximum 200 words)*:

Along with educational, medical, and moral screens, the U.S. Army uses a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new Soldiers. Although the AFQT is useful for selecting new Soldiers, other personal attributes are important to Soldier performance and retention. Based on the U.S. Army Research Institute's (ARI) investigations, the Army selected one promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), for an initial operational test and evaluation (IOT&E), beginning administration to applicants in 2009. Criterion data are being compiled at 6-month intervals from administrative records, from Initial Military Training (IMT), and from schools for eight military occupational specialties (MOS) and will be followed by multiple waves of data collection from Soldiers in units. This is the second of six planned evaluations of the IOT&E. Similar to prior experimental research, our early evaluation suggests that several TAPAS scales significantly predicted a number of criteria of interest, indicating that the measure holds promise for both selection and classification purposes.

15. SUBJECT TERMS

behavioral and social science, personnel, manpower, selection and classification

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	86	Ellen Kinzer Technical Publications Specialist (703) 545-4225

Standard Form 298

Technical Report 1296

Tier One Performance Screen Initial Operational Test and Evaluation: 2010 Annual Report

Deirdre J. Knapp (Ed.)
Human Resources Research Organization

Tonia S. Heffner (Ed.)
U.S. Army Research Institute

Personnel Assessment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

October 2011

Army Project Number
622785A790

Personnel, Performance and Training Technology

ACKNOWLEDGEMENTS

There are individuals not listed as authors who made significant contributions to the research described in this report. First and foremost are the Army cadre who support criterion data collection efforts at the schoolhouses. These noncommissioned officers (NCOs) ensure that trainees are scheduled to take the research measures and provide ratings of their Soldiers' performance in training. Thanks also go to Ms. Kate LaPort, Mr. Irwin Jose, and Ms. Sharon Meyers (ARI) and Mr. Doug Brown, Ms. Ashley Armstrong, and Ms. Mary Adeniyi (HumRRO) and Mr. Jason Vetter (Drasgow Consulting Group) for their contributions to this research effort.

We also want to extend our appreciation to the Army Test Program Advisory Team (ATPAT), a group of senior NCOs who periodically meet with ARI researchers to help guide this work in a manner that ensures its relevance to the Army and help enable the Army support required to implement the research. Members of the ATPAT are listed below:

CSM JOHN R. CALPENA
CSM BRIAN A. HAMM
CSM JAMES SHULTZ
SGM KENAN HARRINGTON
SGM THOMAS KLINGEL
SGM(R) CLIFFORD MCMILLAN
SGM GREGORY A. RICHARDSON
1SG ROBERT FORTENBERRY
MSG JAMES KINSER
MSG ROBERT D. WYATT
SFC WILLIAM HAYES
SFC STEVEN TOSLIN
SFC KENNETH WILLIAMS

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: 2010 ANNUAL REPORT

EXECUTIVE SUMMARY

Research Requirement:

In addition to educational, physical, and moral screens, the U.S. Army relies on a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new Soldiers into the Army. Although the AFQT has proven to be, and will continue to serve as a useful metric for selecting new Soldiers, other personal attributes, in particular non-cognitive attributes (e.g., temperament, interests, and values), are important to entry-level Soldier performance and retention (e.g., Campbell & Knapp, 2001; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009, 2010; Knapp & Tremble, 2007). Based on ARI's research, the Army selected one particularly promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), as the basis for an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen*. TAPAS capitalizes on the latest in testing technology to assess motivation through the measurement of personality characteristics.

In May 2009, the Military Entrance Processing Command (MEPCOM) began administering the TAPAS on the computer adaptive platform for the ASVAB (CAT-ASVAB) at Military Entrance Processing Stations (MEPS). The Work Preferences Assessment (WPA), which asks respondents their preference for various work activities and environments, will be introduced for applicant testing in 2011. Both measures will be administered as part of the IOT&E through 2013. The Information/ Communication Technology Literacy (ICTL) test, developed by the Air Force, also will be administered in 2011 to a subset of applicants as part of the IOT&E. Criterion data are being compiled from administrative records at 6-month intervals. As part of the IOT&E, initial military training (IMT) criterion data are being collected at schools for eight military occupational specialties (MOS) and the first of multiple waves of data collection from Soldiers in their units has been initiated.

Procedure:

Our approach to analyzing the TAPAS' incremental predictive validity was consistent with previous evaluations of this measure and of similar experimental non-cognitive predictors (Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009; 2010; Knapp, Heffner, & White, 2011). In brief, this approach involved testing a series of hierarchical regression models, regressing each criterion measure onto Soldiers' AFQT scores in the first step, followed by their TAPAS scale scores in the second step. When the TAPAS scale scores were added to the baseline regression models, the resulting increment in the multiple correlation (ΔR) served as our index of incremental validity.

As part of the initial TOPS evaluation (Knapp et al., 2011), analyses examined the comparability of the multiple versions of TAPAS on which data are available as well as the classification potential of the measure. These analyses were not repeated for the present report,

but will be revisited in future evaluation reports which are planned for 6-month intervals during the course of the IOT&E.

Findings:

The validation sample sizes were considerably larger for this evaluation compared to the last (e.g., IMT data on 2,297 versus 397 in the last analysis cycle). Results of the selection-oriented analyses suggest that the individual TAPAS scales significantly predict a number of criteria of interest. Most notably, the Physical Conditioning scale predicted Soldiers' self-reported Army Physical Fitness Test (APFT) scores, number of restarts in training, adjustment to Army life, and 3-month attrition. Moreover, the results are consistent with both theoretical descriptions of these scales and previous research (Ingerick et al., 2009; Knapp & Heffner, 2010). However, there is growing evidence of the lack of measurement reliability associated with the performance rating criterion measures which call into question the informativeness of results based on those ratings-based scores. In some cases, the magnitudes of the correlations were smaller than had been found in previous experimental research, however, and the TAPAS composite scores predicted key criteria at a lower rate. Because of the substantive differences between the research and IOT&E contexts, and the preliminary nature of the data, we cannot yet draw a definitive conclusion concerning the reasons for the differences between results from these settings. Several new scales (e.g., Generosity and Adjustment) showed statistically significant correlations with criteria, suggesting that future work should consider updating or revising the selection-oriented composites to enhance the validity of this tool.

The next set of TOPS evaluation analyses will be conducted based on data collected through April and compiled in May 2011. The sample sizes for this next evaluation are expected to be considerably larger, thus supporting additional analyses (e.g., re-examination of the will-do and can-do TAPAS composite scores).

Utilization and Dissemination of Findings:

The research findings will be used by the U.S. Army Accessions Command, U.S. Army Recruiting Command, Army G-1, and Training and Doctrine Command to evaluate the effectiveness of tools used for Army applicant selection and assignment. With each successive set of findings, the Tier One Performance Screen can be revised and refined to meet Army needs and requirements.

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: 2010 ANNUAL REPORT

CONTENTS

Page

Page

Page

List of Tables

CONTENTS (CONTINUED)

Page

Page

List of Figures

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: 2010 ANNUAL REPORT

CHAPTER 1: INTRODUCTION

Deirdre J. Knapp (HumRRO), Tonia S. Heffner and Len White (ARI)

Background

The Personnel Assessment Research Unit (PARU) of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is responsible for conducting personnel research for the Army. The focus of PARU's research is maximizing the potential of the individual Soldier through maximally effective selection, classification, and retention strategies.

In addition to educational, physical, and moral screens, the U.S. Army relies on a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new Soldiers into the Army. Although the AFQT has proven to be and will continue to serve as a useful metric for selecting new Soldiers, other personal attributes, in particular non-cognitive attributes (e.g., temperament, interests, and values), are important contributors to entry-level Soldier performance and retention (e.g., Knapp & Tremble, 2007).

In December 2006, the Department of Defense (DoD) ASVAB review panel—a panel of experts in the measurement of human characteristics and performance— released their recommendations (Drasgow, Embretson, Kyllonen, & Schmitt, 2006). Several of these recommendations focused on supplementing the ASVAB with additional measures for use in selection and classification decisions. The ASVAB review panel further recommended that the use of these measures be validated against performance criteria.

Just prior to release of the ASVAB review panel's findings, ARI initiated a longitudinal research effort, *Validating Future Force Performance Measures (Army Class),* to examine the prediction potential of several non-cognitive measures (e.g., temperament and person-environment fit) for Army outcomes (e.g., performance, attitudes, attrition). The Army Class research project is a 6-year effort that is being conducted with contract support from the Human Resources Research Organization (HumRRO; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009). Experimental predictors were administered to new Soldiers in 2007 and early 2008. Since then, Army Class researchers have obtained attrition data from Army records and collected training criterion data on a subset of the Soldier sample. Job performance criterion data were collected from Soldiers in the Army Class longitudinal validation sample in 2009 (Knapp, Owens, & Allen, 2011) and a second round of job performance data collection was completed in 2011.

After the Army Class research was underway, ARI initiated the *Expanded Enlistment Eligibility Metrics (EEEM)* project (Knapp & Heffner, 2010). The EEEM goals were similar to Army Class, but the focus was specifically on Soldier selection (not classification) and the time horizon was much shorter. Specifically, EEEM required selection of one or more promising new predictor measures for possible immediate implementation. The EEEM project capitalized on the existing Army Class data collection procedure and, thus, the EEEM sample was a subset of the Army Class sample.

As a result of the EEEM findings, Army policy-makers approved an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen (TOPS)*. This report is the second in a series presenting analyses from the IOT&E of TOPS.

The Tier One Performance Screen (TOPS)

Six experimental pre-enlistment measures were included in the EEEM research (Allen, Cheng, Putka, Hunter, & White, 2010). These included several temperament measures, a situational judgment test, and two person-environment fit measures based on values and interests. The "best bet" measures recommended to the Army for implementation were identified based on the following considerations:

- Incremental validity over AFQT for predicting important performance and retention-related outcomes
- Minimal subgroup differences
- Low susceptibility to response distortion (e.g., faking good)
- Minimal administration time requirements

The Tailored Adaptive Personality Assessment System (TAPAS; Stark, Chernyshenko, & Drasgow, 2010b) surfaced as the top choice, with the Work Preferences Assessment (WPA; Putka & Van Iddekinge, 2007) identified as another good option that was substantively different from the TAPAS. TAPAS is a measure of personality characteristics (e.g., achievement, sociability) that capitalizes on the latest in testing technology and provides a good indicator of personal motivation. The WPA asks applicants to indicate their preference for various kinds of work activities and environments (e.g., "A job that requires me to teach others," "A job that requires me to work outdoors"). Although not included in the EEEM research, the Information/Communication Technology Literacy (ICTL) test has emerged as a potential measure of applicants' familiarity with computers and information technology which may predict performance in high-technology occupations.

In May 2009, the Military Entrance Processing Command (MEPCOM) began administering TAPAS on the computer adaptive platform for the ASVAB (CAT-ASVAB). Initially TAPAS was to be administered only to Education Tier 1 (primarily high school diploma graduates), non-prior service applicants. This limitation was removed several months after the start so the Army could evaluate TAPAS across all types of applicants. The TAPAS administration by MEPCOM is scheduled to continue through the fall of 2012.

The Tier One Performance Screen (TOPS) is intended to use non-cognitive measures to identify Education Tier 1 applicants who would likely perform differently (higher or lower) than would be predicted by their ASVAB scores. As part of the TOPS IOT&E, TAPAS scores are being used to screen out a small number of AFQT Category IV applicants.[1] Although the WPA is part of the TOPS IOT&E, it is not under consideration for enlistment eligibility. The WPA is being prepared for MEPS administration with an expected administration start date of fall 2011.

[1] Screening was expanded to include a small number of Category IIIB applicants in July 2011.

Although the initial conceptualization for the IOT&E was to use TAPAS as a tool for "screening in" Education Tier 1 applicants with lower AFQT scores, changing economic conditions spurred a reconceptualization to a system that screens out applicants identified as having low motivation as well as having low AFQT scores.[2] It is likely that the selection model in a fully operational system would adjust to fit with the changing applicant market. For example, at the present time, few applicants are being screened out based on TAPAS scores, not just because the passing scores are set quite low, but also because there are very few Category IV applicants being considered for enlistment due to the overwhelming availability of applicants in higher AFQT categories. Because many factors may impact how TAPAS would be used in the applicant screening process, TAPAS is administered to all Education Tier 1 and Tier 2 non-prior service applicants who take the ASVAB in the MEPS.

Evaluating TOPS

Figure 1.1 illustrates the TOPS IOT&E research plan. To evaluate the non-cognitive measures (TAPAS and WPA), the Army is collecting training criterion data on Soldiers in eight target military occupational specialties (MOS) as they complete initial military training (IMT).[3] The criterion measures include job knowledge tests (JKTs); an attitudinal person-environment fit assessment, the Army Life Questionnaire (ALQ); and performance rating scales (PRS) completed by the Soldiers' cadre. These measures are administered via the Internet at the schools for each of the eight target MOS. The process is overseen by Army personnel with guidance and support from both ARI and HumRRO. Course grades and completion rates are obtained from administrative records for all Soldiers who take the TAPAS, regardless of MOS.

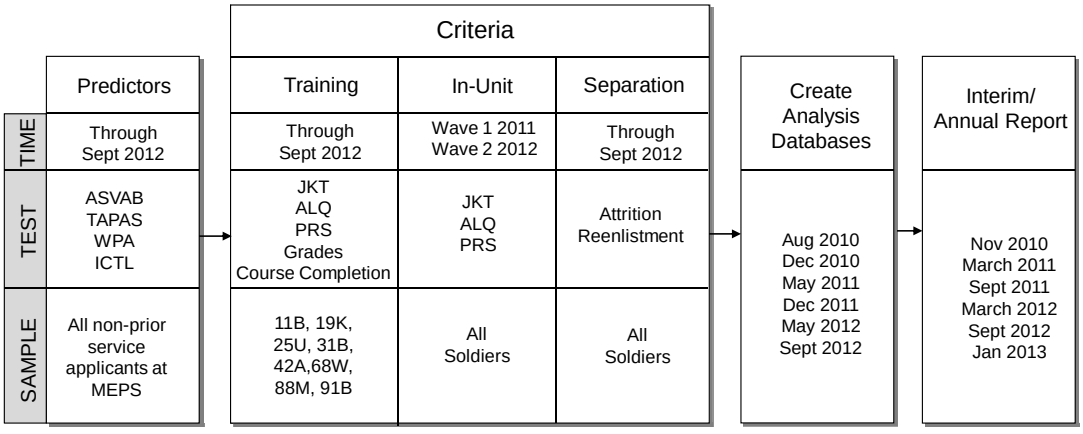

Figure 1.1. TOPS Initial Operational Test & Evaluation (IOT&E).

Two waves of in-unit job performance data collection are also planned, both of which will attempt to capture data from Soldiers from across all MOS who completed the TAPAS during the application process. These measures again will include JKTs, the ALQ, and cadre

[2] Initial supporting data analysis work focused on Category IIIB applicants (Allen et al., 2010), but TOPS currently targets those in Category IV.

[3] The target MOS are Infantryman (11B), Armor Crewman (19K), Signal Support Specialist (25U), Military Police (31B), Human Resources Specialist (42A), Health Care Specialist (68W), Motor Transport Operator (88M), and Light Wheel Vehicle Mechanic (91B).

ratings. Finally, the separation status of all Soldiers who took the TAPAS is being tracked throughout the course of the research.

This report describes the second iteration of an effort to develop a criterion-related validation data file and conduct evaluation analyses using data collected in the TOPS IOT&E initiative. Additional analysis datasets and validation analyses will be prepared and conducted at 6-month intervals throughout the 3-year IOT&E period.

Overview of Report

Chapter 2 explains how the evaluation analysis data files are constructed, then describes characteristics of the samples represented in the latest analysis data file constructed in December 2010. Chapter 3 describes the TAPAS and ASVAB, including content, scoring, and psychometric characteristics. Chapter 4 describes the criterion measures administered to these samples, including their psychometric characteristics. Criterion-related validity analyses are presented in Chapter 5. The report concludes with Chapter 6, which summarizes this second effort to evaluate TOPS (the first was documented in Knapp, Heffner, & White, 2011) and looks toward plans for future iterations of these evaluations.

CHAPTER 2: DATA FILE DEVELOPMENT

D. Matthew Trippe, Laura Ford, Bethany Bynum, and Karen Moriarty (HumRRO)

Overview of Process

The Tier One Performance Screen (TOPS) data file is assembled from a number of sources. In general, the data file comprises predictor and criterion data obtained from administrative and IMT (or "schoolhouse") sources.[4] IMT records comprise assessment data collected from Soldiers and their cadre (i.e., supervisors) at the locations identified in Figure 2.1. The IMT assessments were developed specifically for this IOT&E and are designated below as "non-administrative" to distinguish them from data collected from existing administrative records.

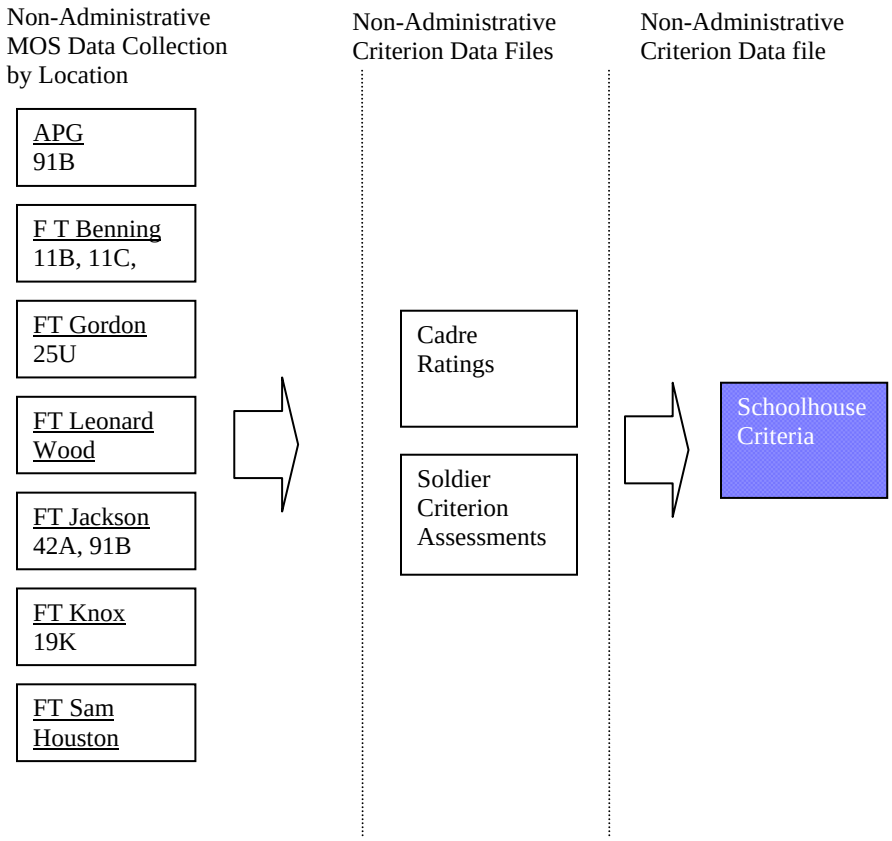

Figure 2.1. Summary of TOPS schoolhouse (IMT) data sources.

A broader view of the entire TOPS analysis file construction process is provided in Figure 2.2. The white boxes within the figure represent source data files, and the shaded boxes represent samples on which descriptive or inferential analyses are conducted. Samples are formed by applying filters to a data file such that it includes the observations of interest. The

[4] Administrative data are collected from the following sources: (a) Military Entrance Processing Command (MEPCOM), (b) Army Human Resources Command (AHRC), (c) U.S. Army Accessions Command (USAAC), and (d) Army Training Support Center (ATSC).

leftmost column in the figure summarizes the predictor data sources used to derive the TOPS Applicant Sample. The other columns summarize the research-only (i.e., non-administrative) and administrative criterion data. Predictor and criterion data are merged to form the schoolhouse-specific validation sample and the full validation sample. The latest version of the TOPS data file (prepared in December 2010) does not contain Work Preferences Assessment (WPA) or Information/Communication Technology Literacy (ICTL) predictor scores or in-unit criteria. Future versions of the data file will be appended with those data.

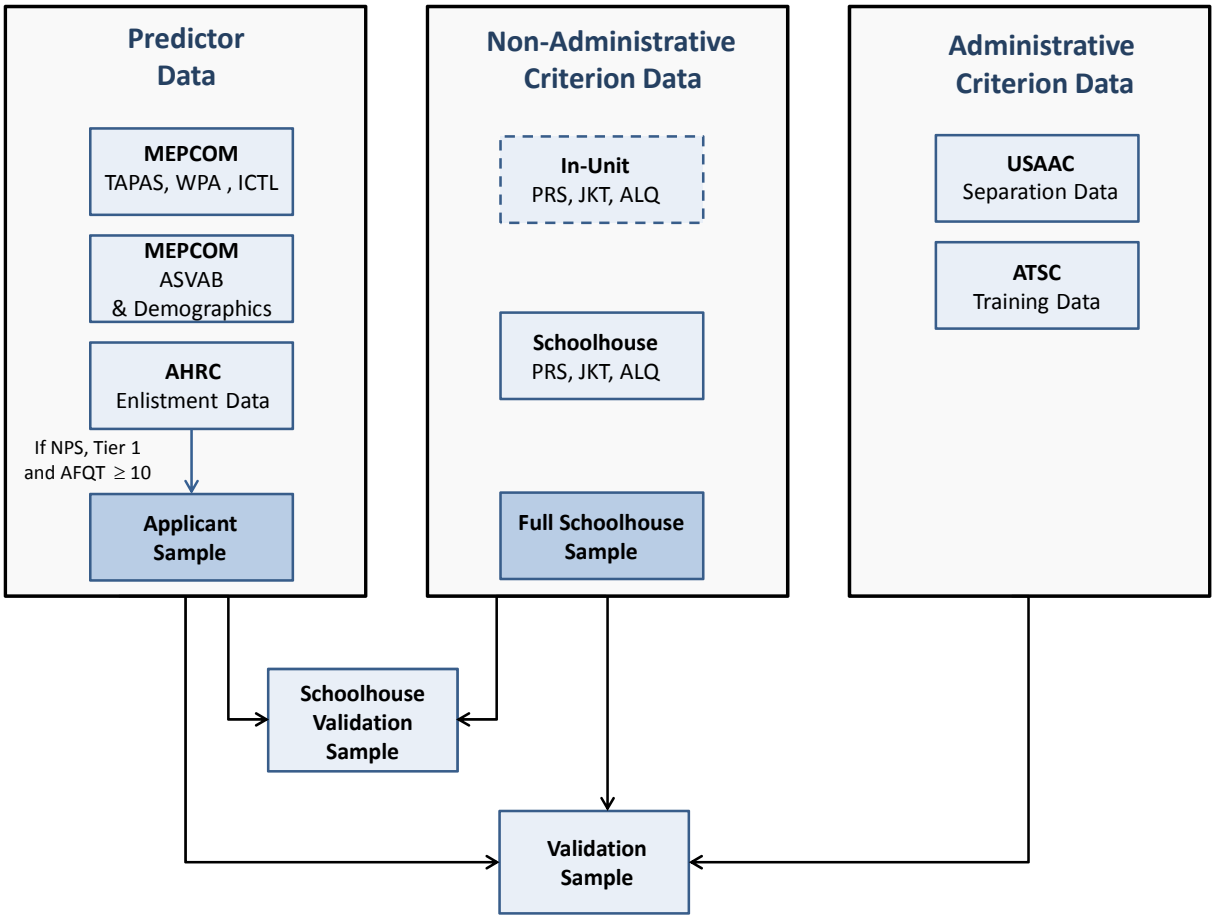

Figure 2.2. Overview of TOPS data file merging and nested sample generation process.

Description of Data File and Sample Construction

Table 2.1 summarizes the total TAPAS sample contained in the December 2010 TOPS data file by key variables that were used to create the samples on which analyses were conducted. The total sample includes applicants who did not necessarily enlist in the Army. The majority of individuals in the data file are classified as Education Tier 1, non-prior service, and AFQT Category I to IV (i.e., AFQT score ≥ 10). [5] All analyses are restricted to these individuals, which results in elimination of approximately 11% of the total records in the data file.

[5] Tier 2 Soldiers were included in the TOPS IOT&E beginning March 2011, after the present data files were constructed.

Table 2.1. Full TAPAS Data File Sample Characteristics

Variables	*n*	% of Total Sample *(N* = 98,331*)*
Education Tier		
Tier 1	92,161	93.7
Tier 2	3,232	3.3
Tier 3	2,938	3.0
Prior Service		
Yes	2,694	2.7
No or Missing	95,637	97.3
MOS		
11B/11C/11X/18X	5,811	5.9
19K	319	0.3
25U	511	0.5
31B	1,332	1.4
42A	586	0.6
68W	1,810	1.8
88M	1,764	1.8
91B	1,443	1.5
Other	18,399	18.7
Unknown[a]	66,356	67.5
AFQT Category		
I	7,799	7.9
II	30,414	30.9
IIIA	19,204	19.5
IIIB	23,769	24.2
IV	15,368	15.6
V	1,775	1.8
Contract Status		
Signed (as of Sept. 2010) [b]	43,996	44.7
Not signed	54,335	55.3
Total Tier 1, Non-prior service (NPS), AFQT ≥ 10 (Applicant Sample)	88,017	89.5

[a]Generally, MOS is unknown either because the respondent did not access into the Army or because the information was not yet available in the data sources on which the December 2010 data file was based.
[b]Signed contract was inferred in 18% of cases where Soldiers had criterion data but no contract date record.

The number and percentage of each MOS represented in the schoolhouse criterion data file is found in Table 2.2. The MOS represented most heavily are11B and 68Ws; least well represented are 19K, 25U, and 42A Soldiers.

Table 2.2. Distribution of MOS in the Full Schoolhouse Sample

MOS	*n*	*%*
11B/11C/11X/18X	7,142	39.7
19K	15	0.1
25U	466	2.6
31B	2,017	11.2
42A	649	3.6
68W	4,338	24.1
88M	2,265	12.6
91B	876	4.9
Other	132	0.7
Unknown	109	0.6
Total	18,009	100.0

A detailed breakout of background and demographic characteristics observed in the analytic samples appears in Table 2.3. Regular Army Soldiers comprise a majority of the cases in each sample. AFQT categories follow an expected distribution. The samples are predominantly male, Caucasian, and non-Hispanic; however a significant percentage of Soldiers declined to provide information on race or ethnicity. The TOPS Applicant Sample was defined by limiting records in the full data file to those Soldiers who are non-prior service, Education Tier 1, and have an AFQT score of at least 10.

The Validation Sample described in Table 2.3 includes 40,944 Soldiers. Those included in this sample are Soldiers who meet all of the inclusion criteria for the TOPS Applicant Sample and also have at least one record in a criterion data source (i.e., Army Training Requirements and Resources System [ATRRS], Resident Individual Training Management System [RITMS], IMT/schoolhouse, attrition). However, the number of Soldiers included in any individual analysis is generally much smaller. The exact number of Soldiers included in a given analysis depends on the criterion variable involved and the limiting factors imposed on that variable (e.g., usability flags, limitations on component). Specific sample details on each criterion variable are provided in subsequent chapters. Generally speaking, 3-month attrition data accounts for approximately 19,000 of these records and the approximately 6,000 administrative graduation and exam records represent the next most available criterion data source.

Although there are 18,009 Soldiers in the full schoolhouse data file, only 2,297 had taken the TAPAS when they applied for enlistment. This disconnect is due largely to the fact that most of the Soldiers tested at the schools had taken their pre-enlistment tests in 2009 before MEPCOM started administering the TAPAS widely to applicants. The problem was exacerbated by the gradual introduction of the TAPAS across MEPS locations so that early in the IOT&E, not all MEPS were yet actively participating. We expect that future analysis data files will show a higher match between Soldiers tested in the schools and those tested pre-enlistment. Indeed, the match rate at this stage (12.7%) is an improvement over the match rate obtained previously (5.5%; Trippe, Ford, Moriarty, & Cheng, 2011).

Table 2.3. Background and Demographic Characteristics of the TOPS Samples

	Applicant[a] *n* = 88,017		Validation[b] *n* = 40,944		Schoolhouse Validation[c] *n* = 2,297	
Characteristic	*n*	*%*	*n*	*%*	*n*	*%*
Component						
Regular Army	56,015	63.6	31,163	76.1	1,436	62.5
Army National Guard	22,763	25.9	6,835	16.7	617	26.9
Army Reserve	9,239	10.5	2,946	7.2	244	10.6
MOS						
11B/11C/11X/18X	5,348	6.1	4,786	11.7	1,066	46.4
19K	305	0.3	281	0.7	1	0.0
25U	475	0.5	308	0.8	9	0.4
31B	1,225	1.4	814	2.0	253	11.0
42A	537	0.6	411	1.0	93	4.0
68W	1,704	1.9	1,313	3.2	474	20.6
88M	1,591	1.8	1,147	2.8	307	13.4
91B	1,285	1.5	983	2.4	94	4.1
Other	17,022	19.3	16,777	41.0	--	--
Unknown	58,525	66.5	14,124	34.5	--	--
AFQT Category						
I	7,301	8.3	3,546	8.7	206	9.0
II	28,257	32.1	14,951	36.5	930	40.5
IIIA	17,569	20.0	9,404	23.0	506	22.0
IIIB	21,310	24.2	11,177	27.3	566	24.6
IV	13,580	15.4	1,866	4.6	89	3.9
V	--	--	--	--	--	--
Gender						
Female	17,241	19.6	6,580	16.1	258	11.2
Male	70,776	80.4	34,364	83.9	2,039	88.8
Race						
African American	12,604	14.3	5,040	12.3	197	8.6
American Indian	603	0.7	287	0.7	18	0.8
Asian	2,464	2.8	1,192	2.9	59	2.6
Hawaiian/Pacific Islander	781	0.9	428	1.0	38	1.7
Caucasian	63,806	72.5	31,658	77.3	1,872	81.5
Multiple	349	0.4	163	0.4	12	0.5
Declined to Answer	7,410	8.4	2,176	5.3	101	4.4
Ethnicity						
Hispanic/Latino	12,650	14.4	5,300	12.9	227	9.9
Not Hispanic	68,169	77.4	33,696	82.3	1,984	86.4
Declined to Answer	7,198	8.2	1,948	4.8	86	3.7

[a] Limited to applicants who had no prior service, Education Tier 1, and AFQT $\geq$ 10.
[b] Limited to applicants who had no prior service, Education Tier 1, and AFQT $\geq$ 10 and had a record in one of the sources used for criterion analyses (i.e., schoolhouse, ATTRS, RITMS, or attrition).
[c] Applicants with schoolhouse data who also had a record in the full TOPS data file.

Summary

The TOPS data file is periodically updated by merging TAPAS scores, administrative records, and IMT data into one master data file. The December 2010 data file includes a total of 98,331 applicants who took the TAPAS, 88,017 of which were in the TOPS Applicant Sample. The Applicant Sample was determined by excluding Education Tier 2, AFQT Category V, and prior service applicants from the master data file. Of that Applicant Sample, 40,944 (47%) had a record in at least one of the criterion data sources and 2,297 (2.6%) had IMT data collected from the schoolhouse. Both of these match rates represent an improvement from the prior reporting cycle. This is likely due to the maturation of criterion data in the source data file files. Higher match rates observed in the present reporting cycle are likely to improve the stability and interpretability of results over the prior cycle. Nevertheless, the amount of criterion data that is actually used in a given analysis remains small in relation to the amount of available predictor data. Subsequent iterations of the TOPS IOT&E data file will no doubt show progressively stronger sample sizes to support validation and other evaluative analyses.

CHAPTER 3: DESCRIPTION OF THE TOPS IOT&E PREDICTOR MEASURES

Stephen Stark, O. Sasha Chernyshenko, Fritz Drasgow (Drasgow Consulting Group), and Matthew T. Allen (HumRRO)

The purpose of this chapter is to describe the predictor measures investigated to date in the TOPS IOT&E (i.e., TAPAS and ASVAB). The central predictor under investigation in this analysis is the Tailored Adaptive Personality Assessment System (TAPAS; Stark, Chernyshenko, & Drasgow, 2010b), while the baseline predictor used by the Army is the ASVAB. We begin this chapter by describing the TAPAS, including previous research and scoring methodology. This is followed by a brief description of the versions administered as part of the TOPS IOT&E. We finish by briefly describing the ASVAB.

Tailored Adaptive Personality Assessment System (TAPAS)

TAPAS Background

TAPAS is a personality measurement tool developed by Drasgow Consulting Group (DCG) under the Army's Small Business Innovation Research (SBIR) program. The system builds on the foundational work of the Assessment of Individual Motivation (AIM; White & Young, 1998) by incorporating features designed to promote resistance to faking and by measuring narrow personality constructs (i.e., facets) that are known to predict outcomes in work settings. Because TAPAS uses item response theory (IRT) methods to construct and score items, it can be administered in multiple formats: (a) as a *nonadaptive test* where examinees respond to the same sequence of items or (b) as an *adaptive* test where each examinee responds to a unique sequence of items selected to maximize measurement accuracy for that specific examinee.

TAPAS uses a recently developed IRT model for multidimensional pairwise preference items (MUPP; Stark, Chernyshenko, & Drasgow, 2005) as the basis for constructing, administering, and scoring personality tests that are designed to reduce response distortion (i.e., faking) and yield normative scores even with tests of high dimensionality (Stark, Chernyshenko, & Drasgow 2010a). TAPAS items consist of pairs of personality statements for which a respondent's task is to choose the one that is "more like me." The two statements composing each item are matched in terms of social desirability and often represent different dimensions. As a result, respondents have a difficult time discerning which answers improve their chances of being enlistment eligible. Because they are less likely to know which dimensions are being used for selection, they are less likely to discern which statements measure those dimensions, and they are less likely to be able to keep track of their answers on several dimensions simultaneously so as to provide consistent patterns of responses across the whole test. Without knowing which answers have an impact on their eligibility status, respondents should not be able to increase their scores on selection dimensions as easily as when traditional, single statement measures are used.

The use of a formal IRT model also greatly increases the flexibility of the assessment process. A variety of test versions can be constructed to measure personality dimensions that are relevant to specific work contexts, and the measures can be administered via paper-and-pencil or computerized formats. If test design specifications are comparable across versions, the respective scores can be readily compared because the metric of the statement parameters has already been

established by calibrating response data obtained from a base or reference group (e.g., Army recruits). The same principle applies to adaptive testing, wherein each examinee receives a different set of items chosen specifically to reduce the error in his or her trait scores at points throughout the exam. Adaptive item selection enhances test security because there is less overlap across examinees in terms of the items presented. Even with constraints governing the repetition and similarity of the psychometric properties of the statements composing TAPAS items, we estimate that over 100,000 possible pairwise preference items can be crafted from the 15-dimension TAPAS pool being used for the IOT&E.

Another important feature of TAPAS is that it contains statements representing potentially 22 narrow personality traits. The TAPAS trait taxonomy was developed using the results of several large scale factor-analytic studies with the goal of identifying a comprehensive set of non-redundant narrow traits. These narrow traits, if necessary or desired, can be combined to form either the Big Five (the most common organization scheme for narrow personality traits) or any other number of broader traits (e.g., Integrity or Positive Core Self-Evaluations). This is advantageous for applied purposes because TAPAS versions can be created to fit a wide range of applications and are not limited to a particular service branch or criterion. Selection of specific TAPAS dimensions can be guided by consulting results of an unpublished meta-analytic study performed by DCG that mapped the 22 TAPAS dimensions to several important organizational criteria for military and civilian jobs (e.g., task proficiency, training performance, attrition).

Three Current Versions of TAPAS

As part of the TOPS IOT&E, three versions of the TAPAS were administered. The first was a 13-dimension computerized adaptive test (CAT) containing 104 pairwise preference items. This version is referred to as the TAPAS-13D-CAT, and was administered from May 4, 2009 to July 10, 2009 to over 2,200 Army and Air Force recruits.[6] In July 2010, ARI decided to expand the TAPAS to 15 dimensions by adding the facets of Adjustment from the Emotional Stability domain and Self Control from the Conscientiousness domain. Test length was also increased to 120 items. Two 15-dimension TAPAS tests were created. One version was nonadaptive (static), so all examinees answered the same sequence of items; the other was adaptive, so each examinee answered items tailored to his or her trait level estimates. The TAPAS-15D-Static was administered from mid-July to mid-September of 2009 to all examinees, and later to smaller numbers of examinees at some MEPS. The adaptive version, referred to as TAPAS-15D-CAT, was introduced in September and Army, Air Force, and Navy recruits continue to be administered this version.[7] Table 3.1 shows the facets assessed by the 13-dimension and 15-dimension measures.

[6] Note that MEPCOM is administering the TAPAS to Air Force applicants on an experimental basis.
[7] Navy recruits began taking the TAPAS 1 April 2011.

Table 3.1. TAPAS Dimensions Assessed

Facet Name	Brief Description	"Big Five" Broad Factor
Dominance	High scoring individuals are domineering, "take charge" and are often referred to by their peers as "natural leaders."	Extraversion
Sociability	High scoring individuals tend to seek out and initiate social interactions.	
Attention Seeking	High scoring individuals tend to engage in behaviors that attract social attention; they are loud, loquacious, entertaining, and even boastful.	
Generosity	High scoring individuals are generous with their time and resources.	Agreeableness
Cooperation	High scoring individuals are trusting, cordial, non-critical, and easy to get along with.	
Achievement	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.	Conscientiousness
Order	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.	
Self Control[a]	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.	
Non-Delinquency	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.	
Adjustment[a]	High scoring individuals are worry free, and handle stress well; low scoring individuals are generally high strung, self-conscious and apprehensive.	Emotional Stability
Even Tempered	High scoring individuals tend to be calm and stable. They don't often exhibit anger, hostility, or aggression.	
Optimism	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.	
Intellectual Efficiency	High scoring individuals are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual.	Openness To Experience
Tolerance	High scoring individuals scoring are interested in other cultures and opinions that may differ from their own. They are willing to adapt to novel environments and situations.	
Physical Conditioning	High scoring individuals routinely participate in vigorous sports or exercise and enjoy physical work.	Other

[a]Not included in TAPAS-13D-CAT.

As part of the first TOPS IOT&E evaluation cycle, descriptive statistics for the TAPAS were computed along with analyses examining the equivalence of these three forms. In general, the results suggested that the three forms were equivalent, and thus could be treated as the same measure provided that the values were standardized within version (Allen, Ingerick, & DeSimone, 2011). With this in mind, the TOPS TAPAS versions were combined into one overall set of TAPAS scales by:

1. Filtering in participants who are part of "Applicant Sample" – Tier 1, non-prior service, AFQT Category IV or above), and

2. Standardizing the variables within version using a *z*-transformation, completed by subtracting each score from the mean for that version and dividing by the standard deviation.

This standardized version of the overall TAPAS was also used in the analyses described in Chapter 5. Readers interested in seeing the details of these analyses should refer to Appendix A.

TAPAS Scoring

TAPAS scoring is based on the MUPP IRT model originally proposed by Stark (2002). The model assumes that when person *j* encounters stimuli *s* and *t* (which, in our case, correspond to two personality statements), the person considers whether to endorse *s* and, independently, considers whether to endorse *t*. This process of independently considering the two stimuli continues until one and only one stimulus is endorsed. A preference judgment can then be represented by the joint outcome (Agree with *s*, Disagree with *t*) or (Disagree with *s*, Agree with *t*). Using a 1 to indicate agreement and a 0 to indicate disagreement, the outcome (1,0) indicates that statement *s* was endorsed but statement *t* was not, leading to the decision that *s* was preferred to statement *t*; an outcome of (0,1) similarly indicates that stimulus *t* was preferred to *s*. Thus, the probability of endorsing a stimulus *s* over a stimulus *t* can be formally written as

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0 \mid \theta_{d_s}, \theta_{d_t}\}}{P_{st}\{1,0 \mid \theta_{d_s}, \theta_{d_t}\} + P_{st}\{0,1 \mid \theta_{d_s}, \theta_{d_t}\}},$$

where:

$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) =$ probability of a respondent preferring statement *s* to statement *t* in item *i*,

$i =$ index for items (i.e., pairings), where $i = 1$ to I,

$d =$ index for dimensions, where $d = 1, \ldots, D$, d_s represents the dimension assessed by statement *s*, and d_t represents the dimension assessed by statement t,

$s, t =$ indices for first and second statements, respectively, in an item,

$(\theta_{d_s}, \theta_{d_t}) =$ latent trait scores for the respondent on dimensions d_s and d_t respectively,

$P_{st}(1,0 \mid \theta_{d_s}, \theta_{d_t}) =$ joint probability of endorsing stimulus *s* and not endorsing stimulus *t* given latent trait scores $(\theta_{d_s}, \theta_{d_t})$,

and

$P_{st}(0,1 \mid \theta_{d_s}, \theta_{d_t}) =$ joint probability of not endorsing stimulus *s* and endorsing stimulus *t* given latent trait scores $(\theta_{d_s}, \theta_{d_t})$.

With the assumption that the two statements are evaluated independently, and with the usual IRT assumption that only θ_{d_s} influences responses to statements on dimension d_s and only θ_{d_t} influences responses to dimension d_t (i.e., local independence), we have

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t})}{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t}) + P_s(0 \mid \theta_{d_s}) P_t(1 \mid \theta_{d_t})},$$

where

$P_s(1 \mid \theta_{d_s}), P_s(0 \mid \theta_{d_s}) =$ probability of endorsing/not endorsing stimulus *s* given the latent trait value θ_{d_s},

and

$P_t(0 \mid \theta_{d_t}), P_t(1 \mid \theta_{d_t}) =$ probability of endorsing/not endorsing stimulus *t* given latent trait θ_{d_t}.

The probability of preferring a particular statement in a pair thus depends on θ_{d_s} and θ_{d_t}, as well as the model chosen to characterize the process for responding to the individual statements. Toward that end, Stark (2002) proposed using the dichotomous case of the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), which has been shown to fit personality data reasonably well (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006).

Test scoring is done via Bayes modal estimation. For a vector of latent trait values, $\tilde{\theta} = (\theta_{d'=1}, \theta_{d'=2}, ..., \theta_{d'=D})$, this involves maximizing:

$$L(\tilde{u}, \tilde{\theta}) = \left\{ \prod_{i=1}^{n} \left[P_{(s>t)_i} \right]^{u_i} \left[1 - P_{(s>t)_i} \right]^{1-u_i} \right\} * f(\tilde{\theta}),$$

where $\tilde{u}$ is a binary response pattern, $P_{(s>t)}$ is the probability of preferring statement *s* to statement *t* in item *i*, and $f(\tilde{\theta})$ is a *D*-dimensional prior density distribution, which, for simplicity, is assumed to be the product of independent normals, $\prod_{d'=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\theta_{d'}^2}{2\sigma^2}}$.

Taking the natural log, for convenience, the above equation can be rewritten as:

$$\ln L(\tilde{u},\tilde{\theta}) = \sum_{i=1}^{n}\left[(u_i)\ln P_{(s>t)_i} + (1-u_i)\ln(1-P_{(s>t)_i})\right] + \sum_{d'=1}^{D}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\theta_{d'}^2}{2\sigma^2}\right] \quad ,$$

leaving the following set of equations to be solved numerically:

$$\frac{\partial \ln L}{\partial \tilde{\theta}} = \begin{bmatrix} \frac{\partial \ln L}{\partial \theta_{d'=1}} \\ \frac{\partial \ln L}{\partial \theta_{d'=2}} \\ \vdots \\ \frac{\partial \ln L}{\partial \theta_{d'=D}} \end{bmatrix} = 0$$

This equation can be solved numerically to obtain a vector of trait score estimates for each respondent via a *D*-dimensional maximization procedure (e.g., Press, Flannery, Teukolsky, & Vetterling, 1990), involving the posterior and its first derivatives. Standard errors for TAPAS trait scores are estimated using a replication method developed by Stark and colleagues (2010a). In brief, this method involves using the IRT parameter estimates for the items that were administered to generate 30 new response patterns based on an examinee's TAPAS trait scores. The resulting simulated response patterns are then scored and the standard deviations of the respective trait estimates over the 30 replications are used as standard errors for the original TAPAS values. In a recent simulation study (Stark, Chernyshenko, & Drasgow, 2010c), this new replication method provided standard error estimates that were much closer to the empirical (true) standard deviations than previously used approaches (i.e., based on the approximated inverse Hessian matrix or a jack-knife approach).

In the present research TAPAS data were flagged as unusable if the applicant selected the same response option more than 63% of the time or, alternatively, if the applicant responded to more than two items in less than 2 seconds each. Descriptive statistics, subgroup differences, and scale intercorrelations for the TAPAS scale scores in the current sample can be found in Appendix B.

TAPAS Initial Validation Effort

Initial predictive and construct-related validity evidence on the TAPAS was collected during ARI's *Expanded Enlistment Eligibility Metrics* (EEEM) research project in 2007-2009 (Knapp & Heffner, 2010). As described in Chapter 1, the EEEM effort was conducted in conjunction with ARI's *Army Class* longitudinal validation of multiple experimental non-cognitive predictor measures. In the EEEM project, new Soldiers completed a 12-dimension, 95-item nonadaptive (or static) version of TAPAS, called TAPAS-95s. TAPAS-95s was administered as a paper questionnaire that included an information sheet showing respondents a sample item and illustrating how to properly record their answers to the "questions" that followed. Respondents were specifically instructed to choose the statement in each pair that was "more like me" and that they must make a

choice even if they found it difficult to do so. Item responses were coded dichotomously and scored using an updated version of Stark's (2002) computer program for MUPP trait estimation.

Overall, the TAPAS-95s showed evidence of construct and criterion validity. Intellectual Efficiency and Curiosity, for example, showed moderate positive correlations with AFQT and correlations of .35 with each other. This was expected, given that both facets tap the intellectance aspects of the Big Five factor, Openness to Experience. The same two traits exhibited similarly positive, but smaller correlations with Tolerance, another facet of Openness reflecting comfortableness around others having different customs, values, or beliefs (Chernyshenko, Stark, Woo, & Conz, 2008). TAPAS-95s dimensions also showed incremental validity over AFQT in predicting several performance criteria. For example, when TAPAS trait scores were added to the regression analysis based on a sample of several hundred Soldiers, the multiple correlation increased by .35 for the prediction of physical fitness, .20 for the prediction of disciplinary incidents, and .11 for the prediction of 6-month attrition. None of these criteria were predicted well by AFQT alone (predictive validity estimates for AFQT were consistently below .10).

The first TOPS IOT&E report expanded on these results by comparing the psychometric properties of the TOPS TAPAS and TAPAS-95s (Knapp et al., 2011). The results of these analyses suggested that (a) the standard deviations for the TOPS TAPAS were, on average, smaller than those found for the TAPAS-95s; (b) some TAPAS scales were more similar across the two settings than others (e.g., Physical Conditioning was consistent, while Attention Seeking was not); and (c) the TOPS TAPAS scales were not strongly related to other individual difference variables (e.g., race, gender), consistent with what was found in EEEM (Allen et al., 2011). Interested readers should again refer to Appendix A for a more in-depth discussion of these results.

In sum, the EEEM research showed TAPAS-95s to be a viable assessment tool with the potential to enhance new Soldier selection. Trait scores exhibited construct validity evidence with respect to other measures and criterion-related validity estimates were fairly high for outcomes not predicted well by AFQT. Based on the results of this research and taking into consideration the unique advantages of TAPAS (e.g., flexibility and resistance to faking), the Army chose to test the measure in an applicant environment. Recognizing that previous research has shown larger differences between experimental and operational use of temperament measures in an Army setting (White, Young, Hunter, & Rumsey, 2008), the results from the first TOPS IOT&E report suggested that the TOPS TAPAS is promising for operational use (Allen et al., 2011). However, certain scales were more consistent psychometrically across settings than others, suggesting that some scales may not maintain the same validity with key criteria of interest.

Initial TAPAS Composites

In addition to the validation analyses described above, an initial Education Tier 1 performance screen was developed from the TAPAS-95s scales for the purpose of testing in an applicant setting (Allen, Cheng, Putka, Hunter, & White, 2010). This was accomplished by (a) identifying key criteria of most interest to the Army, (b) sorting these criteria into "can-do" and "will-do" categories, and (c) selecting composite scales corresponding to the can-do and will-do criteria, taking into account both theoretical rationale and empirical results. The result of this process was two composite scores.

1. Can-Do Composite: The TOPS can-do composite consists of five TAPAS scales and is designed to predict can-do criteria such as MOS-specific job knowledge, AIT exam grades, and graduation from AIT/OSUT.

2. Will-Do Composite: The TOPS will-do composite consists of five TAPAS scales (three of which overlap with the can-do composite) and is designed to predict will-do criteria such as physical fitness, adjustment to Army life, effort, and support for peers.

The target population for these composites was AFQT Category IIIB applicants, though, due to changing recruitment priorities (as described in Chapter 1) the target group was later changed to AFQT Category IV applicants. Specifically, in the TOPS IOT&E, Tier 1 Category IV applicants must score above the 10th percentile on both the can-do and will-do TAPAS

ASVAB Content, Structure, and Scoring

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple aptitude battery of nine tests administered by the Military Entrance Processing Command. Most military applicants take the computer adaptive version of ASVAB (i.e., the CAT-ASVAB). Scores on the ASVAB tests are combined to create composite scores for use in (a) selecting applicants into the Army and (b) classifying them to an MOS. The Armed Forces Qualification Test (AFQT) comprises the Verbal Expression[8] (VE), Arithmetic Reasoning (AR), and Math Knowledge (MK) tests (AFQT = 2*VE + AR + MK). Applicants must meet a minimum AFQT score to be eligible to serve in the military and the Services favor high-scoring applicants for enlistment (e.g., through enlistment bonuses). AFQT percentile scores are divided into the following categories:[9]

- Category I (93-99)
- Category II (65-92)
- Category IIIA (50-64)
- Category IIIB (31-49)
- Category IV (10-30)[10]
- Category V (1-9)

AFQT Category V Soldiers are not eligible for enlistment, while no more than 20% of the total number of enlisted Soldiers can be AFQT Category IV. AFQT Category IIIB applicants are also given lower enlistment priority than AFQT Category I to IIIA applicants.

For classification, scores on the ASVAB tests are combined to form nine Aptitude Area (AA) composites.[11] An applicant must receive a minimum score on the MOS-relevant AA composite(s) to qualify for classification to that MOS. For example, applicants must score a 95 in both the Electronics (EL) and Signal Communications (SC) AA composites to qualify as a Signal Support Specialist (25U). Descriptive statistics for the AFQT, ASVAB tests, and AA composites are reported

[8] Verbal Expression is a scaled combination of the Word Knowledge (WK) and Paragraph Comprehension (PC) tests.
[9] For more information on ASVAB scoring, see the official website of the ASVAB, www.officialasvab.com
[10] AFQT Category IV can be further subdivided into IVA (21-30), IVB (16-20), and IVC (15-15). For the purposes of this report, all AFQT Category IV Soldiers are treated as one group.
[11] A tenth AA composite, General Technical (GT), is not used for enlisted Army selection or classification and therefore is not included here.

in Table B.4 in Appendix B. AFQT Category frequencies are reported in Chapter 2 (Tables 2.1 and 2.3).

Summary

The purpose of this chapter was to describe the predictor measures used as part of the TOPS IOT&E. Three versions of the one experimental measure examined thus far—the TAPAS—were administered as part of the TOPS IOT&E. The TAPAS is unique among typical personality measures because it uses forced-choice pairwise items and IRT to promote resistance to faking. Initial validation research conducted as part of EEEM was promising enough to warrant an IOT&E. Comparative analyses suggest that the three versions of the TAPAS are equivalent, but found some differences with the TAPAS-95s administered as part of EEEM. The validity of both the individual TAPAS scales and can-do and will-do composites formed as part of EEEM are evaluated in Chapter 5. The ASVAB will be used as the baseline instrument for these analyses, which consists of multiple tests that are formed into selection (i.e., AFQT) and classification (i.e., AA) composites.

CHAPTER 4: DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF CRITERION MEASURES

Karen O. Moriarty and Bethany Bynum (HumRRO)

Training criterion measures such as job knowledge tests (JKTs), performance rating scales (PRS), and attitudinal data captured on a self-report questionnaire were used to validate the TAPAS. These measures were originally developed for the training phase of the Army Class project (Moriarty, Campbell, Heffner, & Knapp, 2009), and modified, where needed, for inclusion in the TOPS IOT&E. These measures were used to supplement the administrative data. Table 4.1 summarizes the training criterion measures.

Table 4.1. Summary of Training Criterion Measures

Criterion Measure	Description
Soldier/Cadre Reported	
Job Knowledge Tests (JKT)	MOS-specific JKTs measure Soldiers' knowledge of basic facts, principles, and procedures required of Soldiers in training for a particular MOS. Each JKT includes a mix of item formats (e.g., multiple-choice, multiple-response, and rank order). The Warrior Tasks and Battle Drills (WTBD) JKT measures knowledge that is general to all enlisted Army Soldiers.
Performance Rating Scales (PRS)	PRS measure Soldiers' training performance on two categories: (a) MOS-specific (e.g., learns preventive maintenance checks and services, learns to troubleshoot vehicle and equipment problems) and (b) Army-wide (e.g., exhibits effort, supports peers, demonstrates physical fitness). The PRS are completed by drill sergeants or training cadre.
Army Life Questionnaire (ALQ)	ALQ measures Soldiers' self-reported attitudes and experiences through IMT. The training ALQ focuses on Soldiers' attitudes and experiences in IMT and includes 13 scales that cover (a) Soldiers' commitment and retention-related attitudes, and (b) Soldiers' performance and adjustment.
Administrative	
Attrition	Attrition data were obtained on participating Regular Army Soldiers at 3 and 6 months time-in-service (TIS).
Initial Military Training (IMT) Criteria	These data provide information concerning how many Soldiers restarted IMT and for what reasons, and the number of times Soldiers restarted training.
AIT School Grades	Schoolhouse grades for Soldiers in Advanced Individual Training (AIT).

Training Criterion Measure Descriptions

Job Knowledge Tests (JKTs)

Seven JKTs were developed or adapted for this research: one for Warrior Tasks and Battle Drills (WTBD) which constitute required tasks common for all Soldiers, as well as MOS-specific JKTs for Infantry, Armor, Military Police, Health Care Specialist, Light Wheel Vehicle Mechanic, and Motor Transport Operator. Depending upon the MOS, many JKT items were drawn from items originally developed in prior ARI projects (Campbell & Knapp, 2001; Collins, Le, & Schantz, 2005; Knapp & Campbell, 2006). Most of the JKT items are in a multiple-choice format with two to four response options. However, other formats, such as multiple response (i.e., check all that apply), rank ordering, and matching are also used. The items use visual images to make them more realistic and to reduce reading requirements for the test.

Prior to finalizing the items for use in the TOPS IOT&E, the items were reviewed to ensure they were of high quality. First, we examined the comments Soldiers provided about the assessments during the Army Class testing sessions and made corrections where necessary. For example, several Soldiers in one MOS did not know the meaning of the word, "demarcate," so we changed that word to "mark." Second, we reviewed item statistics from the Army Class data and dropped items that had poor item statistics (e.g., low item-total correlations). Finally, results of the Army Class JKT analyses suggested that the training JKTs were too difficult, so we eliminated the more difficult items.

Performance Rating Scales (PRS)

The PRS also have roots in previous research (see Moriarty et al., 2009 for details). Table 4.2 and Figure 4.1 provide example scales. Depending on the MOS, the number of dimensions ranges from five to nine. The scales were completed by cadre members of the target Soldiers. The scales range from 1 (lowest) to 7 (highest) and include a "not observed" option for instances where the cadre did not have an opportunity to observe a Soldier's performance. They are in the format of a behaviorally-anchored rating scale (BARS), where raters provide one rating per dimension using several examples of high, medium, and low performance as anchors. The scales also include a 3-point "familiarity" rating in which the rater indicates his or her general opportunity to observe each Soldier being rated (i.e., limited, reasonable, or a lot of opportunity to observe).

Table 4.2. Example Training Performance Rating Scales

MOS/AW	Name	Description
Army-Wide	Effort	Puts forth individual effort in study, practice, preparation, and participation activities to complete AIT/OSUT requirements to meet individual Soldier expectations.
MOS-Specific	Learns Safety Procedures	How well has the Soldier learned to follow safety procedures, being alert to possible dangerous or hazardous situations and taking steps to protect self, other Soldiers, and equipment?

A. Overall Performance

Considering your evaluation of the Soldier on the dimensions important to successful performance, please rate the overall effectiveness of each Soldier compared to his/her peers.

1	2	3	4	5
Among the Weakest	**Below Average**	**Average**	**Above Average**	**Among the Best**
(in the bottom 20% of Soldiers)	(in the bottom 40% of Soldiers)	(better than the bottom 40% of Soldiers, but not as good as the top 40%)	(in the top 40% of Solders)	(in the top 20% of Soldiers)

Figure 4.1. Relative overall performance rating scale.

Army Life Questionnaire (ALQ)

The ALQ was designed to measure Soldiers' self-reported attitudes and experiences in training. An earlier form of the ALQ (Van Iddekinge, Putka, & Sager, 2005) was modified slightly for use in the TOPS IOT&E. It focuses on first-term Soldiers' attitudes and experiences in initial military training (IMT) and includes 13 scales that cover (a) Soldiers' commitment and retention-related attitudes, and (b) Soldiers' performance and adjustment. Each ALQ scale is scored differently depending on the nature of the attribute being measured. The Army Physical Fitness Test (APFT) is a write-in item. Training Achievements, Training Failures, and Disciplinary Incidents are simply a sum of the 'YES' responses. The remaining scales (see Table 4.3) use a Likert-type format and are scored by computing a mean of the constituent item scores.

Administrative Criteria

Attrition is a broad category that encompasses involuntary and voluntary separations for a variety of reasons (e.g., underage enlistment, conduct, family concerns, sexual orientation, drugs or alcohol, performance, physical standards or weight, mental disorder, or violations of the Uniformed Code of Military Justice). The reason for separation was determined by the Soldier's Separation Program Designator (SPD) code. Soldiers who were classified as attrits for reasons outside of their or the Army's control were excluded in our analyses (e.g., death or serious injury).

Data on IMT school performance and completion were extracted from the ATRRS and RITMS data files (see Chapter 2). ATRRS course information was used to determine (a) whether a Soldier graduated from or was discharged during IMT and (b) whether he or she restarted during IMT. RITMS data were used to determine Soldiers' Advanced Individual Training (AIT) course grades. Given that each course has different grading procedures, the AIT course grade analysis variable was created by standardizing the grades within course. Due to restricted variance in the One Station Unit Training (OSUT) grades (i.e., all of the grades were pass/fail), these courses were excluded from the course grade analysis variable.

Table 4.3. ALQ Likert-Type Scales

Scale Name	Description	Number of Items	Example Item	Likert Scale Anchors
Affective Commitment	Measures Soldiers' emotional attachments to the Army.	7	I feel like I am part of the Army 'family.'	1 (strongly disagree) to 5 (strongly agree)
Normative Commitment	Measures Soldiers' feelings of obligation toward staying in the Army until the end of their current term of service.	5	I would feel guilty if I left the Army before the end of my current term of service.	1 (strongly disagree) to 5 (strongly agree)
Career Intentions	Measures Soldiers' intentions to re-enlist and to make the Army a career.	3	How likely is it that you will make the Army a career?	Varies by item: 1 (strongly disagree) to 5 (strongly agree); 1 (not at all confident) to 5 (extremely confident); 1 (extremely unlikely to 5 (extremely likely)
Reenlistment Intentions	Measures Soldiers' intention to reenlist in the Army.	4	How likely is it that you will leave the Army after completing your current term of service?	1 (strongly disagree) to 5 (strongly agree)
Attrition Cognition	Measures the degree to which Soldiers think about attriting before the end of their first term.	4	How likely is it that you will complete your current term of service?	Varies by item: 1 (strongly disagree) to 5 (strongly agree); 1 (never) to 5 (very often)
Army Life Adjustment	Measures Soldiers' transition from civilian to Army life	9	Looking back, I was not prepared for the challenges of training in the Army.	1 (strongly disagree) to 5 (strongly agree)
Army Civilian Comparison	Measures Soldiers' impressions of how Army life compares to civilian life.	6	Indicate how you believe conditions in the Army compare to conditions in a civilian job with regards to pay.	1 (much better in the Army) to 5 (much better in civilian life)
MOS Fit	Measures Soldiers' perceived fit with their MOS.	9	My MOS provides the right amount of challenge for me.	1 (strongly disagree) to 5 (strongly agree)
Army Fit	Measures Soldiers' perceived fit with their MOS.	8	The Army is a good match for me.	1 (strongly disagree) to 5 (strongly agree)

Training Criterion Measure Scores and Associated Psychometric Properties

Here we provide a review of the psychometric properties of the training criteria. Basic descriptive statistics are available for the Full Schoolhouse Sample (n = 18,009) and by MOS in Appendix C, along with the intercorrelations. In this chapter we review the psychometric characteristics of the criterion measures estimated using only data on those Soldiers from the TOPS Applicant Sample (i.e., Education Tier 1, non-prior service) whose data were used in the criterion-related validity analyses reported in Chapter 5 (n = 5,968 for schoolhouse IMT criteria,

3,244 for administrative IMT criteria, 9,035 for 3-month attrition, and 4,038 for 6-month attrition). This is referred to as the Validation Sample (see Figure 2.2). Note that the means, standard deviations, and reliability estimates are generally similar to those for the Full Schoolhouse Sample.

Job Knowledge Tests (JKTs)

JKT records were flagged as not useable if the Soldier omitted more than 10% of the assessment items, took fewer than 5 minutes to complete the entire assessment, or chose an implausible response to one of the careless responding items.[12]

A single, overall raw score was computed for each JKT by summing the total number of points Soldiers earned across the JKT items. All of the multiple-choice items were worth one point. Depending on the format of the non-traditional items (e.g., multiple response), they were worth one or more points. To facilitate comparisons across MOS, we computed a percent correct score based on the maximum number of points that could be obtained on each MOS test. For the criterion-related validity analyses, we converted the total raw score to a standardized score (or *z*-score) by standardizing the scores *within* each MOS.

Table 4.4 shows the percent correct scores, as well as internal consistency reliability estimates for the six MOS-specific and the WTBD JKTs. The mean percent correct score across all six MOS-specific tests was 67.0 %, with the 11B and 91B tests being the most difficult (means of 60.30% and 61.31%, respectively). Internal consistency reliability estimates were acceptable, though the WTBD JKT estimate was on the low side (.63) which is not surprising since it covers a broad scope of tasks. Table 4.4 shows the correlations between the various MOS JKT scores with the WTBD JKT score. These are moderate in size and statistically significant. These results suggest that the MOS-specific JKTs and the WTBD JKT each provide some unique performance information.

Table 4.4. Descriptive Statistics and Reliability Estimates for Training Job Knowledge Tests (JKTs) in the TOPS Validation Sample

Test Scores	*n*	*M*	*SD*	*Min*	*Max*	r_{WTBD}	α
11B/11C/11X/18X	852	60.30	9.22	28.26	82.61	**.58**	.77
31B	220	71.59	8.66	45.63	91.26	**.47**	.80
68W	425	75.86	8.73	33.70	92.39	**.48**	.82
88M	247	66.86	10.85	38.89	88.89	**.61**	.77
91B	62	61.31	12.26	29.90	83.51	**.51**	.89
WTBD Job Knowledge	2,176	66.30	12.55	9.68	93.55	--	.63

Note. Mean represents percent correct; α = coefficient alpha. WTBD = Warrior Tasks and Battle Drills. Sample = non-prior service, Education Tier 1, AFQT Category IV or above Soldiers. r_{WTBD} = correlation with WTBD JKT, correlations in bold are statistically significant ($p < .05$). Results for 19K are not reported due to low sample size ($n = 1$).

[12] The 5-minute criterion was established during the first in-unit phase of the Army Class project, which employs highly similar assessments administered via the same platform. See Knapp, Owens, and Allen (2010) for details.

Performance Rating Scales (PRS)

PRS ratings were removed if the cadre member provided a familiarity rating of 1 ("I have had little opportunity to observe this Soldier"). PRS data were also flagged as unusable if the cadre member omitted more than 10% of the assessment items or indicated that he or she did "not observe" the individual on more than 50% of the dimensions. PRS data were also removed if a rater engaged in "flat responding." That is, ratings were removed from the data file if a rater rated 10 or more Soldiers on a particular scale and 90% or more of those rating were exactly the same.

For the MOS-specific PRS, a composite score was created across all of the dimension scores. Computing these scores involved (a) computing the average of multiple ratings provided by cadre (if more than one person rated the target Soldier) and (b) computing the mean of the individual scales that constitute the elements of a particular dimension. The second step was only completed for the MOS-specific PRS, because each of the individual Army-wide scales represented a unique dimension. Approximately 23% of Soldiers in the present sample were rated by more than one cadre member.

Descriptive statistics and estimates of internal consistency reliability for the Army-wide PRS dimensions and MOS PRS composite scores are shown in Table 4.5. Mean ratings are all above average, a common finding in research involving performance ratings. The ratings are also highly correlated (see Appendix C, Table C.2).

Table 4.5. Descriptive Statistics and Reliability Estimates for Training Performance Rating Scales (PRS) in the TOPS Validation Sample

	n	*M*	*SD*	*Min*	*Max*	*α*	*IRR*
Army-Wide Performance Rating Scales							
Effort	726	4.92	1.14	1.00	7.00	n/a	.25
Physical Fitness & Bearing	729	4.94	1.08	1.00	7.00	n/a	.20
Personal Discipline	733	4.98	1.19	1.00	7.00	n/a	.28
Commitment &Adjustment	730	5.07	1.14	1.00	7.00	n/a	.13
Support for Peers	724	5.09	1.10	1.00	7.00	n/a	.17
Peer Leadership	697	4.79	1.21	1.00	7.00	n/a	.18
Common Warrior Tasks Knowledge and Skill	707	4.94	1.10	1.00	7.00	n/a	.11
MOS Qualification Knowledge and Skill	695	4.96	1.10	1.00	7.00	n/a	.15
Overall Performance Scale	726	3.53	0.79	1.00	5.00	n/a	.31
MOS-Specific Performance Rating Composite Scores							
Total (combined across MOS)	532	4.78	0.90	1.00	7.00	n/a	n/a
11B/11C/11X/18X	248	4.98	0.86	2.44	7.00	.94	.12
31B	71	4.80	0.87	2.13	6.50	.94	.37
68W	166	4.38	0.71	1.00	6.20	.91	.00
88M	32	5.41	0.70	4.20	6.80	.90	.01

Note. α = coefficient alpha. n/a = Internal consistency/coefficient alpha could not be computed for the scale. Sample = non-prior service, Education Tier 1, AFQT Category IV or above Soldiers. The possible PRS scores are between 1 and 7, except for the Overall Performance Scale, which ranges from 1 to 5. Results for 19K and 91B are not reported due to low sample sizes (n = 0 and 14, respectively). IRR = Interrater Reliability computed using G(q,k) (Putka, Le, McCloy, & Diaz, 2008).

As illustrated in Table 4.5 and Appendix C, the interrater reliability estimates, which were calculated based on the roughly 23% of the cases for whom there were multiple raters, are quite low. The estimates range from .08 to .24 for the AW scales in the full sample. The highest interrater reliability is associated with the Overall Performance scale on the Army-wide PRS. The low estimates on the MOS-specific rating scales are particularly disturbing given these are composite scores. We attribute these low coefficients to a few interrelated issues. First, the number of ratees per rater is rather high. It averaged 15.5 for the Full Schoolhouse Sample. Second, most raters had very little variance in their ratings, perhaps reflecting their lack of familiarity with individual Soldiers. Third, these data collections were not proctored, while previous studies had administered rating scales such as these in a proctored setting (e.g., Knapp & Heffner, 2009, 2010). Finally, the number of raters per target was small ($k < 2$), which reduces the magnitude of *k*-rater interrater reliability coefficients (see Appendix C). Although not all of these potential issues with the PRS can be addressed within the practical constraints of the research (e.g., collecting ratings in an unproctored setting), the interrater reliability may be improved by revising the PRS format. These revisions are underway.

One strategy to increase the reliability of the ratings of the existing data is to make use of the familiarity ratings in cleaning the data. Accordingly, data from raters indicating that they had little opportunity to observe a Soldier's performance were deleted from the analysis data file. This slightly improved the estimated IRRs and associated validity estimates. Chapter 5 shows the effects of restricting analyses to raters who had a lot of opportunity to observe their Soldiers. Although the effect is positive, it comes at great cost in terms of sample sizes.

In Table C.2 from Appendix C, we see that the correlations among the MOS PRS and the Army-wide PRS in the Full Schoolhouse Sample are moderate to large, with all of them reaching significance. These results suggest there is more content overlap between the MOS PRS and the AW PRS than between the MOS JKTs and WTBD JKT. The AW scale that correlates the strongest with the MOS PRS is, not surprisingly, the MOS Proficiency scale, followed closely by the Common/Warrior Tasks Knowledge and Skills (.69 and .68, respectively).

Army Life Questionnaire (ALQ)

ALQ data were flagged as unusable if the Soldier omitted more than 10% of the assessment items, took fewer than 5 minutes to complete the entire assessment, or chose an implausible response to the careless responding item. ALQ subscale scores were computed in most cases by taking the mean of all responses associated with each scale, properly accounting for reverse coded items. The training failures, training achievement, and disciplinary action scales were computed by summing the total number of "yes" responses.

Table 4.6 provides descriptive statistics and internal consistency reliability estimates for the training ALQ scores. Refer to Table 4.3 for scale anchors, number of items per scale, and sample items. The reliability estimates are good, ranging from .79 to .91. Mean scores are generally similar across MOS (see Table C.3 in Appendix C).

Table 4.6. Descriptive Statistics and Reliability Estimates for the Army Life Questionnaire (ALQ) in the TOPS Validation Sample

Measure/Scale	*n*	*M*	*SD*	*Min*	*Max*	*α*
Affective Commitment	2,232	3.84	0.66	1.00	5.00	.85
Normative Commitment	2,232	4.18	0.68	1.00	5.00	.79
Career Intentions	2,232	3.10	1.06	1.00	5.00	.91
Reenlistment Intentions	2,232	3.41	0.96	1.00	5.00	.84
Attrition Cognition	2,232	1.51	0.60	1.00	5.00	.79
Army Life Adjustment	2,232	4.07	0.65	1.00	5.00	.86
Army Civilian Comparison	2,232	3.89	0.71	1.00	5.00	.80
MOS Fit	2,232	3.83	0.80	1.00	5.00	.91
Army Fit	2,232	4.06	0.58	1.00	5.00	.85
Training Achievement	2,228	0.39	0.59	0.00	2.00	n/a
Training Restarts	2,232	0.36	0.60	0.00	4.00	n/a
Disciplinary Incidents	1,033	0.26	0.59	0.00	6.00	n/a
Last APFT Score	2,209	251.24	31.43	66.00	300.00	n/a

Note. α = coefficient alpha. n/a = Internal consistency/coefficient alpha could not be computed for the scale. Sample = non-prior service, Education Tier 1, AFQT Category IV or above Soldiers. Refer to Table 4.3 for scale anchors, number of items per scale, and sample items.

Administrative Criterion Data

For the first variable, Graduation from IMT, Soldiers who were discharged from the Army during IMT or failed to fully complete their training were coded as 0 (failure). Soldiers who completed IMT and graduated from AIT/OSUT were coded as 1 (graduate). Soldiers who failed to complete their IMT for nonacademic reasons that were administrative in nature and outside the Soldier's control were coded as missing (e.g., returned to unit for mobilization, unit recall, awaiting school start). Soldiers who had not had an opportunity to fully complete their IMT at the time the data were extracted were similarly excluded from our analyses. The second variable, Number of Restarts During IMT, was created by counting the total number of times a Soldier restarted during IMT.

Table 4.7 shows descriptive statistics for the graduation and restart IMT variables. For Soldiers for whom data are available, the attrition rate was 6.8% for 3-month attrition and 10.3% for 6-month attrition. Table D.8 shows that 25U Soldiers had the highest attrition rate (18.2%) and 42A Soldiers had the lowest (.6%). However, note these two had the smallest sample sizes of all the MOS. Overall, 17.6% of the Soldiers restarted at least once during IMT, with the 88M MOS having the largest percentage (22.7%). It is important to note that the IMT data retrieved from administrative sources were not mature for many Soldiers. For example, although there were approximately 44,000 accessed Soldiers in the sample (see Table 2.1), we retrieved attrition data on fewer than 10,000 Soldiers and restart data on fewer than 6,000 Soldiers.

Table 4.7. Descriptive Statistics for Administrative Criteria in the TOPS Validation Sample

Administrative Criterion	N [a]	N_{Attrit}	$\%_{Attrit}$
Attrition			
3-Month Cumulative	9,035	615	6.8
6-Month Cumulative	4,038	417	10.3
Initial Military Training (IMT) Criteria [b]	N [b]	$N_{Restarted}$	$\%_{Restarted}$
Restarted at Least Once During IMT	5,968	715	12.0
Restarted at Least Once During IMT for Pejorative Reasons	5,833	578	9.9
Restarted at Least Once During IMT for Academic Reasons	5,727	474	8.3
AIT School Grades	N [c]	*M*	*SD*
Overall Average (Unstandardized)	3,244	89.83	11.90
Overall Average (Standardized within MOS)	3,244	0.01	0.98

Note. Sample = non-prior service, Education Tier 1, AFQT Category IV or above Soldiers.
[a] N = number of Regular Army Soldiers with 3 or 6-month attrition data at the time data were extracted. N_{Attrit} = number of Soldiers who attrited through 3 or 6 months of service. $\%_{Attrit}$ = percentage of Soldiers who attrited through 3 or 6 months of service [(N_{Attrit} /N) x 100].
[b] N = number of Soldiers with IMT data at the time data were extracted. $N_{Restarted}$ = number of Soldiers who restarted at least once during IMT. $\%_{Restarted}$ = percentage of Soldiers who restarted at least once during IMT [($N_{Restarted}$ /N) x 100].
[c] N = number of Soldiers with AIT school grade data. Standardized school grades were not computed for MOS with insufficient sample size (n < 15).

Summary

Three types of measures were adapted from previous Army research to validate the TAPAS: (a) JKTs, (b) PRS, and (c) the ALQ. Additional criterion data, such as attrition, training restarts, and AIT course grades were gathered from administrative records. The JKTs were completed by Soldiers in eight target MOS and measure MOS-specific and WTBD declarative and procedural knowledge. The PRS were completed by cadre and measure MOS-specific competence and Army-wide constructs such as effort and leadership. Finally, the ALQ asked Soldiers to complete self-report verifiable performance items (e.g., their APFT scores) and attitudinal items (e.g., adjustment to Army life). In general, the criterion measures exhibited acceptable and theoretically consistent psychometric properties. The exception to this was the Army-wide and MOS-specific PRS, which exhibited very low interrater reliability coefficients. Plans to revise the measures and administration procedures in an effort to improve their reliability are discussed in Chapter 6. Until improvements can be implemented, results concerning these scales should be interpreted with caution.

CHAPTER 5: EVIDENCE FOR THE PREDICTIVE VALIDITY AND CLASSIFICATION POTENTIAL OF THE TAPAS

Joseph P. Caramagno, Matthew T. Allen, and Michael J. Ingerick (HumRRO)

This chapter presents the results of analyses examining the operational utility of the TAPAS to improve enlisted Soldier selection. To evaluate the potential of TAPAS to enhance Soldier selection, we begin with measure-level and scale-level results. We then evaluate one potential operational use of the can-do and will-do composites, recognizing that the operational use may change over time. Note that for particular samples of interest (e.g., AFQT Category IV Soldiers) the sample sizes are quite small. Therefore, these results should be treated as preliminary.

Predictive Validity Analyses

Approach

To evaluate the TAPAS' potential to enhance Soldier selection, we examined its incremental validity over the AFQT in predicting early first-term outcomes important to the Army. Consistent with the Army's personnel goals, we selected performance and retention-related outcomes that provided representative coverage of the criterion space. The criterion space for first-term Soldier performance can be specified using three higher-order domains (Campbell, Hanson, & Oppler, 2001; Campbell, McHenry, & Wise, 1990; Strickland, 2005). They are (a) can-do performance, which includes technical and soldiering proficiency; (b) will-do performance, which includes physical, interpersonal, and effort-related criteria; and (c) separation status, which includes attitudes that predict first-term Soldier attrition, and actual attrition behavior. This conceptualization of the performance space resulted in the five criterion groupings shown in Table 5.1, in which can-do performance is addressed by the core technical proficiency measures, retention criteria are covered by another category of scores, and the remaining three categories tap will-do performance. Note that the interrater agreement estimates for the performance rating scales (PRS) reported in the previous chapter were quite low, and therefore the results associated with these scales should be interpreted with caution.[13] Since ratings of Soldiers' overall performance demonstrated the greatest interrater reliability (*IRR* = .31), we included this criterion in the incremental validity analyses for comparison (Table 5.1.).[14]

Our approach to analyzing the TAPAS' incremental predictive validity was consistent with previous evaluations of the measure or similar experimental non-cognitive predictors (Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009; 2010; Trippe, Caramagno, Allen, & Ingerick 2011). In brief, this approach involved testing a series of hierarchical regression models, regressing each criterion measure onto Soldiers' AFQT scores in the first step, followed by their TAPAS scores (scales or composites) in the second step. The resulting increment in the multiple

[13] To address these issues, the PRS are being re-conceptualized (see Chapter 6 for more details).

[14] Overall Performance (PRS) is included separately in Table 5.1 because it encompasses both can-do and will-do performance. As noted in Chapter 4, raters were asked to consider multiple performance dimensions that include cognitive and non-cognitive abilities in making their ratings. Prior evidence (primarily Project A) suggests that, regardless of the dimension being rated (e.g., task knowledge), qualitative ratings capture will-do performance.

correlation (ΔR) when the TAPAS scores were added to the baseline regression models served as our index of incremental validity.

For the continuously scaled criteria, these models were estimated using Ordinary Least Squares (OLS) regression. Alternatively, logistic regression was used for the dichotomous criteria (e.g., 3- and 6-month attrition). At each step in the model, we estimated point-biserial correlations (r_{pb}) in place of the traditional pseudo R estimates to index incremental validity because of conceptual and statistical issues associated with these estimates. The point-biserial correlations reflected the correlation between a Soldiers' predicted probability of engaging in a behavior based on the predictors in the logistic regression model and their actual behavior (e.g., attriting).

In addition to these incremental validity analyses, we examined the predictive validity of the TAPAS at the scale level using bivariate and semi-partial correlations (controlling for AFQT). The semi-partial correlation analyses yield additional information about the individual TAPAS scales' unique contribution to prediction by removing the effects of the AFQT on the TAPAS but not on the criteria (Cohen, Cohen, West, & Aiken, 2003). See Appendix D for the full set of bivariate and semi-partial correlations between the TAPAS composite and scale scores and all of the criteria described in Chapter 4. No corrections for multivariate range restriction or shrinkage were made because of the preliminary nature of these analyses.

Findings

Complete incremental validity analysis results can be found in Appendix D, Table D.1, while a subset of key criteria are presented in Table 5.1. The TAPAS predicted significant incremental variance beyond the AFQT for numerous criteria, especially those related to retention. Consistent with previous research (e.g., Ingerick et al., 2009; Knapp & Heffner, 2009, 2010), the AFQT was generally more predictive of can-do performance-related criteria (Rs ranged from .30 to .49) than will-do performance and retention-related criteria (Rs ranged from .03 to .11). In general, the incremental validity associated with the TAPAS was modest. The largest significant gains were found for Soldiers' self-reported APFT score (ΔR = .21), Army life adjustment (ΔR = .18), cadre ratings of Soldiers' fitness and bearing (ΔR = .17), number of training restarts (ΔR = .16), attrition cognitions (ΔR = .16), and affective commitment to the Army (ΔR = .15). The pattern of ΔRs reported here was similar to that found in the EEEM research. Knapp and Heffner (2010) reported incremental validities for three of the criteria shown in Table 5.1, including MOS-Specific JKTs (ΔR = .03), Last APFT Score (ΔR = .26), and Affective Commitment (ΔR = .18). The relationship between the AFQT and these matched performance- and retention-related criteria in TOPS was also comparable to that observed in EEEM (MOS-Specific JKT, R = .48, $p < .05$; Last APFT Score, R = .04, *ns*; Affective Commitment, R = .06, *ns*), though the AFQT was more predictive of Last APFT Score in the current sample.

Raters were asked to gauge their familiarity with the Soldiers they were rating using a 3-point scale. As discussed in Chapter 4, we hypothesize that some of the low reliability associated with the ratings may be due to the relative lack of opportunity for raters to observe relevant behaviors. Previously reported analyses for the TOPS IOT&E (Trippe et al., 2011) and the EEEM project (Allen, Cheng, Putka, Hunter, & White, 2010) did not use the familiarity ratings

to remove cases from the analysis sample. In the present analyses, we have removed ratings from cadre who indicated that they had had "little opportunity to observe this Soldier" (i.e., a familiarity rating of 1). This data cleaning step did increase the IRR estimates and strength of results (see Table 5.1).

Table 5.1. Incremental Validity Estimates for the TAPAS Scales over the AFQT for Predicting Select Performance- and Retention-Related Criteria

Criterion	*n*	AFQT Only *R* (r_{pb})	AFQT + TAPAS *R* (r_{pb})	Δ*R* (Δr_{pb})
Can-Do Performance				
Core Technical Proficiency				
WTBD Job Knowledge Test (WTBD JKT)	1,992	**.493**	**.499**	.005
MOS-Specific JKT	1,661	**.378**	**.392**	.014
IMT Exam Grade	2,920	**.299**	**.313**	**.014**
Will-Do Performance				
Effort and Leadership				
Effort (PRS)	649	.046	.195	.149
Training Restarts (ALQ)	2,040	**.084**	**.243**	**.159**
Peer Leadership (PRS)	621	.028	.168	.141
Maintaining Personal Discipline				
Personal Discipline (PRS)	654	**.089**	**.223**	**.133**
Disciplinary Action (ALQ)	943	.044	.149	.104
Physical Fitness and Military Bearing				
Physical Fitness and Bearing (PRS)	650	.037	**.205**	**.168**
Last APFT Score (ALQ)	2,018	**.108**	**.320**	**.211**
Retention				
Affective Commitment (ALQ)	2,040	**.090**	**.235**	**.145**
Adjustment to Army Life (ALQ)	2,040	**.086**	**.262**	**.175**
Attrition 3-Months	8,242	**(.058)**	**(.105)**	**(.047)**
Attrition 6-Months	3,439	**(.062)**	**(.123)**	**(.061)**
Overall Performance (PRS)	649	**.084**	**.212**	.128
PRS Criteria with Familiarity = "A Lot"				
Effort and Leadership				
Effort (PRS)	339	**.111**	**.296**	**.185**
Peer Leadership (PRS)	329	**.111**	.271	.160
Maintaining Personal Discipline				
Personal Discipline (PRS)	339	**.164**	**.315**	**.151**
Physical Fitness and Military Bearing				
Physical Fitness and Bearing (PRS)	337	.075	**.285**	**.209**
Overall Performance (PRS)	339	**.174**	**.295**	.121

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. PRS = Performance Rating Scales. AFQT Only = Correlation between the AFQT and the criterion of interest. AFQT + TAPAS = Multiple correlation (*R*) between the AFQT and the selected predictor measure with the criterion of interest. Δ*R* = Increment in *R* over the AFQT from adding the selected predictor measure to the regression model ([AFQT + TAPAS] – AFQT Only). Estimates in parentheses are *point-biserial correlations (r_{pb})* that reflect the observed point-biserial correlation between Soldiers' predicted probability of attriting and their actual attrition behavior. Large, positive r_{pb} values mean that the TOPS composite or scale performed well in predicting actual attrition. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed). The PRS data in the top half of the table include supervisors with familiarity ratings of 2 ("I have had reasonable opportunity to observe this Soldier") and 3 ("I have had a lot of opportunity to observe this Soldier"), while the bottom half include supervisors with familiarity ratings of 3 only,

Though it would decrease sample sizes more dramatically, we could go even further and limit the analysis to raters that had "a lot of opportunity to observe" the Soldiers that they are rating (i.e., a familiarity rating of 3). To demonstrate the effect of familiarity on incremental validity, Table 5.1 also displays incremental validity coefficients for the TAPAS scales over the AFQT for ratings that are limited to raters with the highest level of self-assessed familiarity. Sample sizes decreased by roughly half and the ΔRs increased by less than .05, but did increase systematically. The overall pattern of results did not change; however, the incremental validity estimate for ratings of effort became significant when the ratings were limited to supervisors with the highest degree of familiarity with the target Soldiers.

Note that the smallest correlations and validity estimates are associated with the PRS, which may simply reflect criterion unreliability. TAPAS provided a small, significant increment in the prediction of ratings of discipline ($\Delta R = .13$, $p < .05$) and physical fitness ($\Delta R = .17$, $p < .05$), which is consistent with prior research (Knapp & Heffner, 2010), but it did not significantly predict ratings of effort, leadership, peer support, MOS qualification and skill, or MOS-specific performance over the AFQT alone. The TAPAS did not demonstrate significant incremental validity for Overall Performance ($\Delta R = .13$, *ns*), but it did approach significance ($p = .053$) and the TAPAS had the highest IRR estimates which may explain why this scale showed at least a nominal improvement in its observed relationship to TAPAS. In sum, the accuracy of conclusions we can draw based on the PRS criteria at this time is limited.

Table 5.2 displays the bivariate and semi-partial correlations between the scores on the TAPAS scales and composites and the selected criterion measures. There were a number of notable statistically significant relationships that were consistent with a theoretical understanding of the TAPAS scales and previous research (Knapp & Heffner, 2010). For example, Achievement and Dominance were positively correlated with Training Achievement, last APFT score, and numerous retention-related criteria such as Adjustment to Army Life, Affective Commitment, and Army/MOS Fit (*r*'s ranged from .05 to .15). Larger correlations were found for Physical Conditioning which was positively correlated with self-reported APFT score ($r = .27$) and Adjustment to Army Life ($r = .18$), and negatively correlated with Number of Restarts ($r = -.18$). Intellectual Efficiency was positively correlated with WTBD JKT ($r = .23$), MOS-specific JKT ($r = .14$), IMT Exam Grade ($r = .12$), and Adjustment to Army Life ($r = .12$). A number of other TAPAS scales, including Achievement, Adjustment, Attention Seeking, Dominance, and Optimism, significantly predicted Adjustment to Army Life and Affective Commitment. Optimism also significantly predicted 3- and 6-month attrition. There were a few other statistically significant correlations, such as Generosity, Order, and Sociability being negatively correlated with WTBD and MOS-specific Job Knowledge and Sociability being negatively correlated with IMT Exam Grade. The TAPAS composites generally performed as expected. The can-do composite correlated positively with job knowledge criteria and IMT Exam Grade, and the will-do composite correlated positively with Last APFT Score and negatively with Disciplinary Incidents.

Examination of the scale-level incremental validity coefficients in Table 5.2 shows that this general pattern of results remained largely the same after controlling for AFQT, suggesting the TAPAS' impact on the criteria of interest is largely independent of AFQT. The notable exception was for Intellectual Efficiency, whose correlations with can-do performance criteria

(WTBD JKT, MOS JKT, IMT Exam Grade) dropped to nearly zero after controlling for AFQT. This finding makes theoretical sense and is consistent with prior research in which Intellectual Efficiency has emerged as the TAPAS scale most strongly correlated with the AFQT (Knapp & Heffner, 2010). Not surprisingly, correlations between the can-do composite (which contains the Intellectual Efficiency scale) and the previously mentioned can-do performance criteria also dropped to near-zero and became non-significant after controlling for AFQT. In summary, the overall pattern of results suggests that the relationships between the TAPAS scales and the criteria are generally independent of AFQT, with a few exceptions.

Similar to the incremental validity analyses, Table 5.3 displays correlations between select PRS criteria and the TAPAS dimensions and composites by level of supervisors' familiarity. Limiting the analyses to supervisors that report having "a lot of opportunity" (as opposed to "reasonable opportunity") to observe the Soldiers they rated reduced the sample sizes by roughly half and the increase in the magnitude of the estimates was generally negligible (mean $|\Delta r| = .02$). While most of the significant estimates remained unchanged, the correlations between two of the TAPAS dimensions (Attention Seeking and Order) and Personal Discipline became non-significant in the constrained sample and the correlation between Even Tempered and Effort increased by .07 and became significant.

Table 5.2. Bivariate and Semi-Partial Correlations between the TAPAS Scales and Selected Criteria

	Criteria																	
	Can-Do Performance								Will-Do Performance				Retention					
	WTBD JKT		MOS-Specific JKT		IMT Exam Grade		Training Restart (ALQ)		Disciplinary Incidents (ALQ)		Last APFT Score (ALQ)		Adjustment to Army Life (ALQ)		Affective Commitment (ALQ)		3-Month Attrition[b]		6-Month Attrition[b]	
TAPAS Dimensions	*n* = 2,100		*n* = 1,751		*n* = 3,098		*n* = 2,151		*n* = 996		*n* = 2,128		*n* = 2,151		*n* = 2,151		*n* = 8,638		*n* = 3,810	
Achievement	**.05**	(.00)	.01	(-.03)	.03	(.00)	**-.09**	**(-.08)**	**-.10**	**(-.09)**	**.08**	**(.06)**	**.13**	**(.12)**	**.12**	**(.13)**	.00	(.01)	.00	(.00)
Adjustment[a]	**.08**	(.03)	.03	(-.01)	.00	**(-.04)**	**-.05**	(-.04)	-.02	(-.02)	.01	(.00)	**.09**	**(.08)**	**-.05**	(-.04)	-.02	(-.01)	-.01	(.00)
Attention Seeking	.03	(-.02)	.03	(-.01)	-.03	**(-.06)**	-.03	(-.02)	.03	(.04)	.03	(.02)	**.05**	**(.04)**	**.07**	**(.08)**	-.01	(-.01)	**-.04**	(-.03)
Cooperation	-.02	(-.02)	.01	(.01)	-.03	(-.03)	**.04**	**(.04)**	.01	(.01)	-.02	(-.02)	-.03	(-.03)	.00	(.00)	-.01	(-.01)	.00	(.00)
Dominance	.03	(-.01)	-.03	**(-.07)**	.01	(-.02)	**-.09**	**(-.08)**	-.04	(-.03)	**.13**	**(.12)**	**.15**	**(.14)**	**.12**	**(.13)**	.00	(.01)	-.02	**(-.01)**
Even Tempered	.02	(-.02)	.03	(.00)	.02	(-.01)	.01	(.02)	.03	(.03)	**-.08**	**(-.09)**	.03	(.03)	.00	(.01)	-.01	(-.01)	.00	(.00)
Generosity	**-.05**	(-.01)	**-.05**	(-.03)	-.03	(-.01)	**.04**	(.04)	-.06	(-.06)	.01	(.02)	.00	(.00)	**.08**	**(.07)**	**.04**	(.03)	.02	(.02)
Intellectual Efficiency	**.23**	(.02)	**.14**	(-.02)	**.12**	(-.01)	**-.10**	**(-.07)**	-.02	(.00)	.04	(-.01)	**.12**	**(.10)**	-.02	(.02)	-.01	**(.02)**	-.01	**(.02)**
Non-delinquency	-.03	(-.03)	-.01	(-.01)	.03	**(.04)**	**.06**	**(.06)**	-.01	(-.01)	**-.08**	**(-.08)**	.02	(.02)	**.05**	**(.05)**	.01	(.01)	.02	(.02)
Optimism	.00	(.00)	.01	(.01)	-.02	(-.02)	-.01	(-.01)	-.01	(-.01)	.02	(.02)	**.12**	**(.11)**	**.08**	**(.08)**	**-.03**	(-.03)	**-.04**	**(-.04)**
Order	**-.08**	(.00)	**-.11**	(-.04)	-.01	**(.04)**	.00	(-.01)	-.01	(-.02)	.03	**(.05)**	.03	**(.04)**	.01	(-.01)	**.02**	**(.01)**	.03	(.02)
Physical Conditioning	.03	(.00)	-.01	(-.03)	.00	(-.02)	**-.18**	**(-.18)**	-.06	(-.05)	**.27**	**(.27)**	**.18**	**(.17)**	**.06**	**(.06)**	**-.05**	(-.05)	**-.07**	**(-.07)**
Self Control[a]	-.01	(-.01)	-.06	**(-.05)**	.01	(.01)	.01	(.01)	.01	(.01)	-.02	(-.02)	.02	(.03)	**.05**	**(.04)**	.00	**(.00)**	.01	(.01)
Sociability	**-.07**	(-.03)	**-.07**	(-.04)	**-.06**	**(-.04)**	.01	(.01)	.02	(.01)	.02	(.03)	.02	(.03)	.03	(.02)	.00	(-.01)	-.02	(-.02)
Tolerance	-.04	(-.03)	-.02	(-.01)	**-.04**	(-.03)	**.08**	**(.08)**	-.02	(-.02)	.02	(.02)	.03	(.03)	.04	(.04)	.00	(.00)	.02	(.02)
TAPAS Composites																				
Can-Do	**.09**	(-.02)	**.06**	(-.02)	**.07**	(.00)	**-.04**	(-.03)	-.04	(-.03)	-.01	(-.03)	**.15**	**(.14)**	**.09**	**(.11)**	-.01	(.00)	-.02	(.00)
Will-Do	.00	(-.02)	-.01	(-.03)	**.05**	(.03)	**-.07**	**(-.06)**	**-.07**	**(-.06)**	**.06**	**(.06)**	**.13**	**(.12)**	**.07**	**(.07)**	-.01	(-.01)	-.01	(-.01)

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in parentheses are semi-partial correlations between the TAPAS scales and the criterion of interest, controlling for AFQT. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 478 – 8,242.

[b] Attrition results include Regular Army Soldiers only.

Table 5.3. Bivariate and Semi-Partial Correlations between the TAPAS Scales and Composites and Select Performance Rating Scale Criteria by Supervisor Familiarity Level

	Criteria							
	Familiarity = "Reasonable" and "A Lot"				Familiarity = "A Lot"			
TAPAS Dimensions	Effort *n* = 696	Peer Leadership *n* = 668	Personal Discipline *n* = 703	Physical Fitness and Bearing *n* = 699	Effort *n* = 358	Peer Leadership *n* = 348	Personal Discipline *n* = 358	Physical Fitness and Bearing *n* = 356
Achievement	.05 (.05)	.05 (.05)	.06 (.05)	.04 (.04)	.07 (.06)	.04 (.03)	.06 (.05)	.02 (.02)
Adjustment[a]	-.03 (-.03)	-.05 (-.05)	-.04 (-.05)	-.05 (-.05)	-.04 (-.06)	-.06 (-.07)	-.08 (-.10)	-.05 (-.05)
Attention Seeking	**.08 (.08)**	.06 (.06)	**.08** (.07)	**.09 (.08)**	**.12 (.10)**	.10 (.09)	.06 (.05)	**.11** (.10)
Cooperation	.02 (.02)	-.04 (-.04)	-.07 (-.07)	-.01 (-.01)	.04 (.04)	-.09 (-.09)	-.09 (-.09)	.01 (.01)
Dominance	.06 (.06)	.02 (.02)	.02 (.01)	.06 (.06)	.04 (.03)	.03 (.03)	.02 (.00)	.07 (.06)
Even Tempered	.05 (.05)	.00 (.00)	.05 (.04)	.04 (.04)	**.12 (.11)**	.00 (.00)	.08 (.07)	.08 (.07)
Generosity	-.05 (-.04)	-.02 (-.01)	-.03 (-.02)	-.03 (-.02)	-.03 (-.02)	.00 (.00)	.00 (.02)	.05 (.05)
Intellectual Efficiency	-.01 (-.04)	.01 (.00)	.01 (-.03)	-.04 (-.06)	.00 (-.05)	.03 (-.02)	.02 (-.06)	-.03 (-.07)
Non-delinquency	.01 (.01)	.03 (.03)	-.01 (-.01)	-.03 (-.03)	-.08 (-.08)	-.02 (-.02)	-.08 (-.08)	-.09 (-.09)
Optimism	.03 (.03)	.05 (.05)	.05 (.04)	.04 (.04)	.03 (.03)	.02 (.02)	.05 (.05)	.06 (.06)
Order	.06 (.07)	**.10 (.11)**	**.09 (.11)**	.06 (.07)	.07 (.09)	**.12 (.15)**	.10 **(.13)**	.03 (.04)
Physical Conditioning	**.11 (.11)**	.05 (.04)	.02 (.02)	**.11 (.11)**	**.13 (.12)**	.08 (.08)	.05 (.04)	**.17 (.17)**
Self Control[a]	.01 (.02)	-.02 (-.02)	.03 (.03)	-.01 (-.01)	.00 (.00)	-.08 (-.07)	-.03 (-.03)	-.02 (-.02)
Sociability	.00 (.00)	.00 (.01)	-.05 (-.05)	.01 (.01)	.02 (.03)	.06 (.07)	-.01 (.00)	.07 (.08)
Tolerance	.01 (.01)	-.01 (-.01)	.01 (.02)	.02 (.02)	-.01 (-.01)	.00 (.00)	-.02 (-.02)	.00 (.00)
TAPAS Composites								
Can-Do Composite	.04 (.03)	.05 **(.05)**	.05 (.03)	.02 (.01)	.04 (.02)	.02 (.00)	.04 (.00)	.02 (.00)
Will-Do Composite	.05 (.05)	.03 (.03)	.01 (.01)	.03 (.03)	.04 (.04)	-.01 (-.01)	.01 (.00)	.03 (.03)

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in parentheses are semi-partial correlations between the TAPAS scales and the criterion of interest, controlling for AFQT. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 621 – 654 ("reasonable" and "a lot") and 329 – 339 ("a lot" only).

Finally, we computed correlations between the TAPAS composite scores and the selected criteria by AFQT category to explore the potential influence this might have on our results (see Table 5.4). There were few statistically significant results, and sample sizes varied substantially across the AFQT categories.[15] The TAPAS can-do composite demonstrated significant, positive correlations with Adjustment to Army Life across all AFQT categories and Affective Commitment for Category II, IIIA, and IIIB scores. The will-do composite correlated significantly and positively with Adjustment to Army Life for Category I, II, and IIIA Soldiers and self-reported APFT scores for Category II and IIIA Soldiers. Neither composite was significantly correlated with 3- or 6-month across all AFQT categories. Though larger, generally non-significant estimates were found for Category IV Soldiers, sample sizes were smallest in this group (*n* ranged from 35 – 357) and the results might not reflect true performance on these criteria for the "typical" Category IV Soldier. In general, correlations tended to be small ($|r| < .20$) with many at or near zero.

Evaluation of Initial TAPAS Can-Do and Will-Do Screens

Approach

To evaluate the current can-do and will-do screens, we conducted split-group analyses comparing Soldiers' criterion scores by AFQT category (i.e., I – IV) and TOPS pass/fail status, with particular attention to the cutoff points differentiating Category IIIA from Category IIIB Soldiers and Category IIIB from Category IV Soldiers. For select continuously-scaled criteria, we computed effect sizes (Cohen's *d*) for Category IIIB and IV Soldiers that pass TOPS and those that fail TOPS, then compared these groups to the next highest AFQT category as well as to each other within category. For the purposes of this analysis, the minimum passing score for the TAPAS composites was set as the 10th percentile within AFQT Categories IIIB and IV. Specifically, Soldiers scoring in the 10th percentile or lower within AFQT Categories IIIB and IV on *either* the TOPS can-do or will-do composite were considered to "fail" the TOPS screen, while those above the 10th percentile on both composites received a "passing" score on the TOPS screen.

To compare the groups of interest, we used a Cohen's *d* for the continuously-scaled criteria and a Relative Risk (RR) ratio for the dichotomously-scaled criteria. Cohen's *d* is a measure of effect size comparing the means for a given criterion between two groups. A smaller *d* statistic indicates greater similarity between the groups. The RR ratio is a proportion comparing the percentage of Soldiers that had an incident (i.e., attrited) to the percentage that did not have an incident (i.e., did not attrit).

[15] For these criteria, sample sizes ranged from 9 to 3,116. The smallest sample sizes were generally associated with Category IV Soldiers' scores on criteria related to performance rating scales (PRS) (e.g., Peer Leadership). Estimates with sample sizes less than 30 were omitted from the analyses.

Table 5.4. Correlations between TAPAS Composite Scores and Select Performance and Retention-Related Criteria

	AFQT Category									
	Cat I		Cat II		Cat IIIA		Cat IIIB		Cat IV	
TAPAS Composite/Criterion	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*
TAPAS Can-Do Composite										
Core Technical Proficiency										
WTBD Job Knowledge Test	196	.03	865	.02	460	**.10**	496	**-.09**	83	-.16
MOS-Specific JKT	174	.11	759	-.02	382	.00	381	-.04	55	-.17
IMT Exam Grade	258	.02	1,109	-.04	705	.03	876	**.07**	150	-.04
Effort and Leadership										
Effort (PRS)	82	-.06	316	.02	152	.09	135	.08	--	--
Training Restarts (ALQ)	200	-.05	885	-.03	473	-.04	509	.01	84	-.13
Peer Leadership (PRS)	79	-.01	306	.02	148	.07	126	.10	--	--
Maintaining Personal Discipline										
Personal Discipline (PRS)	82	-.13	320	.05	155	.02	135	.08	--	--
Disciplinary Action (ALQ)	94	-.18	373	-.04	233	.03	261	-.06	35	.16
Physical Fitness and Military Bearing										
Physical Fitness and Bearing (PRS)	81	-.17	318	.05	154	.06	135	-.05	--	--
Last APFT Score (ALQ)	200	.04	877	-.02	466	.01	502	-.06	83	-.20
Retention										
Affective Commitment (ALQ)	200	.07	885	**.08**	473	**.13**	509	**.12**	84	.20
Adjustment to Army Life (ALQ)	200	**.15**	885	**.17**	473	**.10**	509	**.10**	84	**.24**
Attrition 3-Months	762	.06	3,116	.00	1,958	-.04	2,445	.00	357	.03
Attrition 6-Months	385	.08	1,422	-.01	857	-.03	991	.01	155	-.09
TAPAS Will-Do Composite										
Core Technical Proficiency										
WTBD Job Knowledge Test	196	.07	865	.00	460	.03	496	-.07	83	-.06
MOS-Specific JKT	174	.06	759	-.04	382	.01	381	-.02	55	-.07
IMT Exam Grade	258	.06	1,109	-.01	705	**.10**	876	.04	150	.13
Effort and Leadership										
Effort (PRS)	82	-.01	316	.05	152	.12	135	.03	--	--
Training Restarts (ALQ)	200	-.07	885	**-.08**	473	-.05	509	-.06	84	-.09
Peer Leadership (PRS)	79	-.10	306	.04	148	.04	126	.07	--	--
Maintaining Personal Discipline										
Personal Discipline (PRS)	82	-.11	320	.06	155	-.03	135	-.01	--	--
Disciplinary Action (ALQ)	94	-.13	373	-.08	233	-.01	261	-.09	35	.07
Physical Fitness and Military Bearing										
Physical Fitness and Bearing (PRS)	81	-.08	318	.07	154	.06	135	-.07	--	--
Last APFT Score (ALQ)	200	.13	877	**.07**	466	**.11**	502	.03	83	-.18
Retention										
Affective Commitment (ALQ)	200	.00	885	**.07**	473	.08	509	.07	84	.15
Adjustment to Army Life (ALQ)	200	**.14**	885	**.16**	473	**.09**	509	.09	84	.18
Attrition 3-Months	762	.03	3,116	-.01	1,958	-.02	2,445	-.02	357	.05
Attrition 6-Months	385	.03	1,422	.00	857	-.04	991	.00	155	-.06

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed). Cells with sample sizes less than 30 were omitted from the analyses.

Findings

Results of the split-group analyses are reported in Tables 5.5 and 5.6. Some results with smaller sample sizes (i.e., $n < 15$) were omitted from the analyses. As expected, Category IIIA Soldiers tended to have higher WTBD JKT scores, MOS-specific JKT scores, IMT Exam Grades, and Affective Commitment to the Army, as well as lower rates of attrition, higher training pass rates, and fewer disciplinary incidents than Category IIIB Soldiers (Table 5.5). Category IIIB Soldiers that failed TOPS generally demonstrated lower criterion scores, higher risk for attrition, and more disciplinary incidents than those that passed TOPS. The exceptions to this trend were found in four criteria – WTBD JKT, Peer Leadership, Personal Discipline, and Training Restarts – where Category IIIB Soldiers that failed TOPS performed better than Category IIIB Soldiers that passed TOPS. The lack of reliability for the Peer Leadership and Personal Discipline performance ratings, however, make these findings questionable. Nonetheless, none of the *d*-coefficients for the comparisons between Category IIIB Soldiers that passed and those that failed TOPS were statistically significant.

Results for the comparison of Category IIIB and IV Soldiers (Table 5.6) were mixed. The TAPAS composites seemed to enhance the AFQT in discriminating between higher and lower potential Soldiers on the retention-related criteria (e.g., Affective Commitment). However, Category IV Soldiers that failed TOPS performed better than those that passed TOPS on a comprehensive job knowledge criterion. Due to limited sample sizes, we were not able to compute Relative Risk ratios between Category IV Soldiers that pass and fail TOPS.

Summary

Overall, these analyses provide an assessment of the operational deployment of the TAPAS to enhance new Soldier selection. The TAPAS scales enhanced the AFQT in predicting a variety of behaviors considered important to the Army, especially those related to retention such as adjustment to Army life, perceptions of fit with the Army, and affective commitment. In addition, scores on the TAPAS scales generally predicted will-do performance criteria at a higher rate than can-do performance beyond the AFQT. Specifically, the TAPAS demonstrated small to moderate incremental validity over the AFQT for predicting Soldiers' physical fitness and bearing and discipline. Analysis of bivariate and semi-partial correlations between the TAPAS and criteria of interest provided evidence that the gain in predictive potential was largely independent from the AFQT. Finally, the current TAPAS composites demonstrated partial utility in identifying low potential candidates to "select out" of the Army. However, in this sample some high potential Category IIIB and IV Soldiers that performed as well as or better than their peers would have been screened out. This was particularly true of AFQT Category IV Soldiers.

These results are likely due, in part, to the composition of the current TAPAS composites, which include three individual TAPAS scales that are statistically unrelated to most of the can-do and will-do performance criteria (i.e., bivariate correlations at or near zero.)[16] However, as mentioned previously, the low reliability estimates for the PRS criteria diminish the likelihood that TAPAS scores will correlate with performance ratings. This might, in fact, have a

[16] It should be noted that Achievement correlates ($p < .05$) with Training Achievement and Training Restarts in the expected direction, however the magnitudes of the correlations are below |.10|.

greater impact on the results because the TAPAS was specifically designed to predict will-do performance. That being said, two issues support the conclusion that the next round of analyses should reconsider the can-do and will-do TAPAS composites. First, the composites were developed using a smaller sample in an experimental setting. The current applicant sample might not compare in terms of the added strain on the Soldiers whose scores on the TAPAS impact actual selection into the Army (i.e., the current sample might take the assessment more seriously). Second, the current, expanded set of TAPAS scales includes four that were not available during the initial development of the composites (i.e., Adjustment, Generosity, Self Control, and Sociability). A reassessment of the TAPAS composites could result in replacing the poorer performing scales with one or more of these alternate scales.

Table 5.5. AFQT Category IIIB Split Group Analysis Comparing Soldiers by AFQT Category on Targeted Continuous and Dichotomous Criteria

	AFQT Category – Continuous Criteria										
	I-II		IIIA		IIIB Pass TOPS		IIIB Pass-IIIA	IIIB Fail TOPS		IIIB Fail-IIIA	IIIB Fail-IIIB Pass
Criterion	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*d*	*d*
Performance Rating Scales	*(n = 397-414)*		*(n = 154-162)*		*(n = 102-109)*			*(n = 24-26)*			
Effort	4.97	1.13	4.91	1.12	4.86	1.20	-0.04	4.70	1.22	-0.19	-0.14
Peer Leadership	4.81	1.19	4.80	1.18	4.65	1.34	-0.13	4.81	1.24	0.01	0.12
Personal Discipline	5.08	1.15	4.90	1.19	4.73	1.28	-0.14	4.82	1.29	-0.07	0.07
Self-Report Criteria	*(n = 953-1,112)*		*(n = 394-493)*		*(n = 328-437)*			*(n = 53-72)*			
WTBD JKT	0.47	0.84	-0.10	0.93	-0.53	0.99	**-0.46**	-0.42	0.84	**-0.34**	0.12
MOS-Specific JKT	0.38	0.91	-0.01	0.97	-0.39	0.93	**-0.39**	-0.44	0.82	**-0.44**	-0.05
Last APFT Score	254.41	30.32	248.48	31.93	247.94	33.13	-0.02	246.94	25.98	-0.05	-0.03
Affective Commitment	3.80	0.68	3.82	0.67	3.94	0.58	**0.18**	3.79	0.65	-0.05	-0.26
Adjustment to Army Life	4.14	0.60	3.97	0.73	4.03	0.64	0.09	3.98	0.62	0.02	-0.08
Administrative Criteria	*(n = 1,407)*		*(n = 729)*		*(n = 750)*			*(n = 126)*			
IMT Exam Grade	0.27	0.89	-0.01	0.97	-0.26	0.99	**-0.26**	-0.39	1.03	**-0.40**	-0.14

	AFQT Category – Dichotomous Criteria										
	I-II		IIIA		IIIB Pass TOPS		IIIB Pass-IIIA	IIIB Fail TOPS		IIIB Fail-IIIA	IIIB Fail-IIIB Pass
	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*N*	*RR*	$\%_{INCDNT}$	*N*	*RR*	*RR*
Training Restarts	10.8	271	11.0	158	14.2	196	1.29	11.9	24	1.08	0.84
Disciplinary Incidents	20.0	96	19.8	48	22.8	50	1.15	--	--	--	--
3-Month Attrition	5.3	213	7.1	145	8.8	186	1.25	10.5	36	1.48	1.19

Note. IIIB Pass/Fail-IIIA d = standardized mean difference in criterion scores (or Cohen's *d*) between the selected IIIB and IIIA Soldiers [(M_{IIIB} - M_{IIIA}) / SD_{IIIA}]; coefficients in bold are significant at $p < .05$ using an independent samples *t*-test. *IIIB Fail-IIIB Pass d* = standardized mean difference in criterion scores (or Cohen's *d*) between the selected IIIB Soldiers [($M_{IIIB\ Fail}$ - $M_{IIIB\ Pass}$) / $SD_{IIIB\ Pass}$]; coefficients in bold are significant at $p < .05$ using an independent samples *t*-test. $\%_{INCDNT}$ = % of Soldiers, out of the total number (*N*), in selected AFQT Category that exhibited an incident on the criterion measure (e.g., attriting within 3 months). *IIIB-IIIA RR* = relative risk ratio of selected IIIB Soldiers having an incident relative to that of IIIA Soldiers (*p*[IIIB having an incident]/*p*[IIIA having an incident]); coefficients in bold are statistically significant at $p < .05$ using a Chi-square test. Sample sizes less than 15 were omitted from these analyses.

Table 5.6. AFQT Category IV Split Group Analysis Comparing Soldiers by AFQT Category on Targeted Continuous and Dichotomous Criteria

	AFQT Category – Continuous Criteria										
	I-IIIA		IIIB		IV Pass TOPS		IV Pass -IIIB	IV Fail TOPS		IV Fail-IIIB	IV Fail-IV Pass
Criterion	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*d*	*d*
Performance Rating Scales	*(n = 551-576)*		*(n = 137-146)*		*(n = 7-8)*			*(n = 2-3)*			
Effort	4.95	1.13	4.82	1.18	--	--	--	--	--	--	--
Peer Leadership	4.81	1.19	4.69	1.31	--	--	--	--	--	--	--
Personal Discipline	5.03	1.16	4.75	1.29	--	--	--	--	--	--	--
Self-Report Criteria	*(n = 1,347-1,605)*		*(n = 403-539)*		*(n = 45-67)*			*(n = 10-17)*			
WTBD JKT	0.30	0.91	-0.53	0.96	-0.93	0.95	**-0.42**	-0.78	1.02	-0.26	0.16
MOS-Specific JKT	0.27	0.95	-0.40	0.92	-0.50	0.78	-0.11	--	--	--	--
Last APFT Score	252.60	30.93	247.74	31.79	247.33	35.20	-0.01	251.18	40.03	0.11	0.11
Affective Commitment	3.81	0.68	3.91	0.61	4.09	0.58	**0.30**	3.72	0.56	-0.30	**-0.64**
Adjustment to Army Life	4.08	0.65	4.01	0.64	4.19	0.56	**0.28**	3.87	0.57	-0.22	**-0.58**
Administrative Criteria	*(n = 2,136)*		*(n = 947)*		*(n = 138)*			*(n = 12)*			
IMT Exam Grade	0.18	0.93	-0.29	1.00	-0.48	1.06	-0.19	--	--	--	--

	AFQT Category – Dichotomous Criteria										
	I-IIIA		IIIB		IV Pass TOPS		IV Pass -IIIB	IV Fail TOPS		IV Fail-IIIB	IV Fail-IV Pass
	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*N*	*RR*	$\%_{INCDNT}$	*N*	*RR*	*RR*
Training Restarts	10.9	429	13.9	240	13.2	33	0.95	--	--	--	--
Disciplinary Incident	19.9	144	22.5	62	--	--	--	--	--	--	--
3-Month Attrition	5.9	358	8.9	233	5.9	19	0.66	--	--	--	--

Note. IV Pass/Fail-IIIB d = standardized mean difference in criterion scores (or Cohen's *d*) between the selected IV and IIIB Soldiers [(M_{IV} - M_{IIIB}) / SD_{IIIB}]; coefficients in bold are significant at $p < .05$ using an independent samples t-test. *IV Fail-IV Pass d* = standardized mean difference in criterion scores (or Cohen's *d*) between the selected IV Soldiers [($M_{IV\ Fail}$ - $M_{IV\ Pass}$) / $SD_{IV\ Pass}$]; coefficients in bold are significant at $p < .05$ using an independent samples t-test. $\%_{INCDNT}$ = % of Soldiers, out of the total number (*N*), in selected AFQT Category that exhibited an incident on the criterion measure (e.g., attriting within 3 months). *IV-IIIB RR* = relative risk ratio of selected IV Soldiers having an incident relative to that of IIIB Soldiers (*p*[IV having an incident]/*p*[IIIB having an incident]); coefficients in bold are statistically significant at $p < .05$ using a Chi-square test. Sample sizes less than 15 were omitted from these analyses.

CHAPTER 6: SUMMARY AND A LOOK AHEAD

Deirdre J. Knapp (HumRRO), Tonia S. Heffner and Leonard A. White (ARI)

Summary of the TOPS IOT&E Method

The Army is conducting an initial operational test and evaluation (IOT&E) of the Tier One Performance Screen (TOPS). The TOPS assessments, including the Tailored Adaptive Personality Assessment Screen (TAPAS), the Information/Communication Technology Literacy (ICTL) test, and soon the Work Preferences Assessment (WPA), are being administered to non-prior service applicants testing at MEPS locations.

To evaluate the TAPAS, ICTL, and WPA, the Army is collecting training criterion data on Soldiers in selected MOS as they complete their Initial Military Training (IMT). The criterion measures include job knowledge tests (JKTs); an attitudinal person-environment fit assessment, the Army Life Questionnaire (ALQ); and performance rating scales (PRS) completed by the Soldiers' cadre members. Course grades and completion rates are obtained from administrative records for all Soldiers, regardless of MOS. The plan is to construct analysis datasets and conduct validation analyses at 6-month intervals throughout the IOT&E period.

At least two waves of in-unit job performance data collection are also planned at approximately 18 month intervals, each attempting to capture data from Soldiers from across all MOS who completed the TAPAS, and WPA and ICTL as available, at entry. These measures will again include JKTs, the ALQ, and supervisor ratings. Finally, the separation status of all Soldiers who took the TAPAS at entry is being tracked throughout the course of the research.

Summary of Evaluation Results to Date

A staggered schedule for getting schoolhouse testing underway along with the fact that there is generally an appreciable delay between when individuals take pre-enlistment tests and when they access into the Army resulted in low match rates between TAPAS scores and IMT data, but the overall validation sample sizes are reasonable. Of greater concern at this stage is the low reliability associated with the performance rating measures which limits the extent to which the true relationships between the constructs they are intended to measure and predictor measures can be demonstrated. Thus, the analyses conducted thus far must be viewed with some caution. Following is a brief summary of the analyses conducted over the course of the first two TOPS evaluation cycles.

TAPAS Construct Validity

The three versions of the TAPAS (13D-CAT, 15D-Static, and 15D-CAT) are consistent with one another in terms of their means, standard deviations, and patterns of intercorrelations. The two computer-adaptive versions of the TAPAS are particularly similar. Some of the TAPAS scales appear more similar across research and operational settings than others. The patterns of relations between TAPAS scales and individual difference variables (AFQT scores, race, ethnicity, and gender), however, were generally consistent from the EEEM to TOPS settings.

Keeping in mind that previous research has shown large differences between the experimental and operational use of temperament measures (White et al., 2008), these results suggest that the use of the TAPAS in an operational setting is promising.

Validity for Soldier Selection

The results of the selection-oriented analyses suggest that the individual TAPAS scales significantly predict a number of criteria of interest. In addition, many of these correlations are theoretically consistent with expectations. Most notably, the Physical Conditioning scale predicted Soldiers' self-reported APFT scores, number of restarts, adjustment to Army life, affective commitment, and both 3- and 6-month attrition. The Optimism scale also significantly predicted affective commitment as well as attrition. Intellectual Efficiency predicted scores on the job knowledge tests, initial military training (IMT) Exam Grades, and number of training restarts. A number of scales (Achievement, Adjustment, Intellectual Efficiency, Physical Conditioning, and Optimism) predicted the Adjustment to Army Life scale. These results are consistent with both theoretical descriptions of these scales and previous research (Ingerick et al., 2009; Knapp & Heffner, 2010), supporting the use of these scales in an operational setting. Given that some of the scales are not included in either the can-do or will-do composites (e.g., Adjustment), but did predict aspects of Soldier performance, future work will develop more comprehensive selection-oriented composites.

Potential for Soldier Classification

In the initial evaluation cycle, Trippe, Caramagno, Allen, and Ingerick (2011) examined the classification potential of the TAPAS by looking at MOS differences in TAPAS score profiles. Mean differences (evaluated by computing the overall average root mean squared difference in scale scores) for the overall TAPAS were comparatively smaller than those observed in the ASVAB. The magnitude of the differences varied by TAPAS scale, however, often in ways that are consistent with a theoretical understanding of the scale and the MOS. For example, the means for Physical Conditioning were higher for more physically-oriented MOS, such as 11B and 31B. The mean for the Intellectual Efficiency scale was highest for 68W, the most cognitively-oriented MOS in the sample. Results examining the predictive validity estimates found that the Adjustment, Intellectual Efficiency, and Optimism scales generally exhibited the largest differences across MOS.

Taken together, these early evaluation results suggest that, while the magnitude of the validity and classification coefficients are not as large as those found in the experimental EEEM research (Knapp & Heffner, 2010), the TAPAS holds promise for both selection and classification-oriented purposes. Many of the scale-level coefficients are consistent with a theoretical understanding of the TAPAS scales, suggesting that the scales are measuring the characteristics that they are intended to measure. However, given the restricted nature of the matched criterion sample (in terms of sample characteristics) and the low reliability of the ratings data, these results should be considered highly preliminary. Future analyses will expand on these findings by examining operational applications of TAPAS, such as developing new selection and classification composites and determining the effect of various cut scores.

Looking Ahead

Predictor Measures

Three new versions of TAPAS will be introduced into the MEPS. Each of the 15 dimension versions will have nine core dimensions that are consistent across versions and include all of the scales in the "can do" and "will do" composites. Six dimensions, which were included in the original version of TAPAS and have shown promise for initial entry selection, are included on two of the three TAPAS version. Six new scales are being tested and evaluated on a single TAPAS version (see Table 6.1). The dimensions will be evaluated for potential use as core dimensions on later versions of TAPAS. The current version of TAPAS will continue to be used in the research environment.

Table 6.1. TAPAS Dimensions Assessed

	Version A	Version B	Version C
Achievement	✓	✓	✓
Adjustment	✓	✓	✓
Adventure Seeking		✓	
Attention Seeking	✓	✓	✓
Commitment to Serve		✓	
Cooperation	✓	✓	
Courage			✓
Dominance	✓	✓	✓
Even Tempered	✓	✓	✓
Intellectual Efficiency	✓	✓	✓
Non-Delinquency	✓	✓	✓
Optimism	✓	✓	✓
Order	✓	✓	
Physical Conditioning	✓	✓	✓
Responsibility			✓
Self Control	✓		✓
Selflessness	✓	✓	
Sociability	✓		✓
Situational Awareness		✓	
Team Orientation			✓
Tolerance	✓		✓

In 2011, to support Air Force initiatives, the ICTL was administered to 52,000 Army, Air Force and Navy applicants. For the more than 25,000 Army applicants, the ICTL will be examined as an additional predictor of performance, attitudes, and attrition. It also is anticipated that MEPCOM will begin administering the WPA to Army applicants in 2011.

Criterion Measures

In the spring of 2011, the MOS-specific and WBTD JKTs (both training and in-unit versions) are being reviewed and updated with the assistance of Army subject matter experts. As part of this effort, additional items will be added to the WBTD JKT in an effort to increase both its reliability and content representativeness. Additional items are being added to the 31B JKTs

to cover content domains that have increased in relevance since the test blueprint was originally developed. In addition to updating and improving existing measures, we are developing MOS-specific measures (both training and in-unit) for two occupations – Signal Support Specialist (25U) and Human Resources Specialist (42A).

Experience thus far with the performance rating scales suggests rethinking our approach to these measures. Plans for revising both the training and in-unit performance rating scales (which we expect would have similar, though less severe psychometric problems) are under discussion. For example, we are changing the format of the training MOS-specific rating scales to use a 5-point relative performance rating rather than a 7-point absolute performance rating and to greatly reduce the amount of reading required. The training Army-wide PRS will be similarly changed, and the number of dimensions rated will be reduced. Final decisions regarding these changes and their implementation into the TOPS IOT&E will be discussed further in the next reporting cycle.

In-Unit Data Collections

Collection of data from Soldiers in units who took the TAPAS prior to enlistment began in April 2011. The data collection model closely mirrors that which was used in the Army Class research program (Knapp, Owens, & Allen, 2010). To the extent possible, we will visit major Army installations and Reserve Component training sites to collect Soldier and supervisor data in proctored settings. Other Soldiers will provide data from self-administered testing sessions. Regardless of setting, all measures will be administered via the Internet.

Analyses

The semi-annual reports will include basic psychometric, validation, and incremental validation analyses. As more data become available, we will conduct analyses to develop and evaluate new TAPAS composite scores. We will examine the comparability of the new TAPAS versions to prior forms before determining if the data can be combined for purposes of analysis. Analysis strategies also will be developed to handle data produced by substantially revised performance rating scales starting later in 2011. This will be a particular challenge in the training validation sample and may require truncation of some future analyses to include only data provided by the newer measures to provide the best criterion-related validity evidence. Finally, the plan is to conduct classification-oriented analyses once annually.

The analyses reported thus far have been restricted to Education Tier 1 Soldiers (high school degree graduates) because (a) they are the focus of the original TOPS concept and (b) this allows relatively direct comparison of these results to those obtained in a more purely research setting (i.e., the *Expanded Enlistment Eligibility Metrics* project). Because the Army has expanded TAPAS testing to Tier 2, we may consider alternative selection models, just as the composite scores will get re-examined. Future evaluations might include Soldiers in other Education Tiers.

A third set of TOPS evaluation analyses will be conducted based on a data file constructed in May 2011. The sample sizes for this next evaluation are expected to be considerably larger, thus supporting additional analyses (e.g., re-examination of how the will-do and can-do TAPAS composite scores are constructed) and yielding more generalizable results.

REFERENCES

Allen, M.T., Cheng, Y.A., Putka, D.J., Hunter, A., & White L. (2010). Analysis and findings. In D.J. Knapp & T.S. Heffner (Eds.). *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Allen, M.T., Ingerick, M.J., & DeSimone, J.A. (2011). Psychometric evaluation of the TAPAS. In D.J. Knapp, T.S. Heffner, & L. White (Eds.) *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J.P., Hanson, M. A., & Oppler S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* Hillsdale, NJ: Erlbaum.

Campbell, J.P., & Knapp, D.J. (Eds.) (2001). *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43,* 313-333.

Chernyshenko, O.S., Stark, S., Drasgow, F., & Roberts, B.W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88-106.

Chernyshenko, O.S., Stark, S., Woo, S., & Conz, G. (2008, April). *Openness to Experience: Its facet structure, measurement and validity*. Paper presented at at the 23nd annual conference for the Society of Industrial and Organizational Psychologists. New Orleans, LA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd edition.* Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S., (2003). *Applied multiple regression/ correlation analysis for the behavioral and social sciences* (3rd ed.), Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, M., Le, H., & Schantz, L. (2005). Job knowledge criterion tests. In D.J. Knapp & T.R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Technical Report 1168) (pp. 49-58). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Drasgow, F., Embretson, S.E., Kyllonen, P.C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Ingerick, M., Diaz, T., & Putka, D. (2009). *Investigations into Army enlisted classification systems: Concurrent validation report* (Technical Report 1244). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., & Campbell, R.C. (Eds). (2006). *Army enlisted personnel competency assessment program: Phase II report* (Technical Report 1174). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., & Heffner, T.S. (Eds.) (2009). *Predicting Future Force Performance (Army Class): End of Training Longitudinal Validation* (Technical Report 1257). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Heffner, T. S. (Eds.). (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., Heffner, T.S., & White, L. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., & Owens, K.S., & Allen, M.T. (Eds.) (2011). *Validating Future Force Performance Measures (Army Class): First In-Unit Performance Longitudinal Validation* (Technical Report 1293). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., & Tremble, T.R. (Eds) (2007). *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McMichael, W. H. (2008, October 14). Shaky economy helps recruiting, retention. *Army Times*. Retrieved November 3, 2010, from http://www.armytimes.com/news/2008/10 /military_recruiting_2008_101008w/.

McMichael, W. H. (2009, October 15). Economy fueled recruiting gains in FY09. *Army Times*. Retrieved November 3, 2010, from http://www.armytimes.com/news/2009/10/ military_recruiting_retention_101309w/.

Moriarty, K.O., Campbell, R.C., Heffner, T.S., & Knapp, D.J. (2009). *Validating future force performance measures (Army Class): Reclassification test and criterion development* (Research Product 2009-11). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1990). *Numerical recipes: The art of scientific computing*. New York: Cambridge University Press.

Putka, D. J., Le, H., McCloy, R. A., Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*, 959-981.

Putka, D.J., & Van Iddekinge, C.H. (2007). Work Preferences Survey. In D.J. Knapp & T.R. Tremble (Eds.), *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Schafer, S. M. (2007, May 4). Good economy makes recruiting tough, personnel chief says. *Associated Press*. http://www.armytimes.com/news/2007/05/ap_economyrecruit_070504/.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment* [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Stark, S.E., Chernyshenko, O.S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise preference model. *Applied Psychological Measurement, 29*, 184-201.

Stark, S.E., & Chernyshenko, O.S., & Drasgow, F. (2010a). Adaptive testing with the Multi-Unidimensional Pairwise Preference model. Manuscript submitted for publication.

Stark, S.E., Chernyshenko, O.S., & Drasgow, F. (2010b). Tailored adaptive personality assessment system (TAPAS-95s). In D.J. Knapp & T.S. Heffner (Eds.) *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Stark, S.E., Chernyshenko, O.S., & Drasgow, F. (September, 2010c).*Update on the Tailored Adaptive Personality Assessment System (TAPAS): Results and ideas to meet the challenges of high stakes testing.* Paper presented at the 52nd annual conference of the International Military Testing Association. Lucerne, Switzerland.

Stark, S.E., Chernyshenko, O.S., & Drasgow, F., & Williams, B.A. (2006). Item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.

Stark, S. E., Hulin, C. L., Drasgow, F., & Lopez-Rivas, G. (2006). *Technical report 2 for the SBIR Phase II funding round, topic A04-029: Behavior domains assessed by TAPAS* (DCG200608). Urbana, IL: Drasgow Consulting Group.

Strickland, W.J. (Ed.) (2005). *A longitudinal examination of first term attrition and reenlistment among FY1999 enlisted accessions* (Technical Report 1172). Alexandria, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Trippe, D.M, Caramagno, J.P., Allen, M.T., & Ingerick, M.J. (2011). Initial evidence for the predictive validity and classification potential of the TAPAS. In D.J. Knapp, T.S. Heffner, & L. White (Eds.) *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Trippe, D.M., Ford, F., Moriarty, K.O., & Cheng, Y.A. (2011). Database development. In D.J. Knapp & T.R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (pp. 89-104) (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Van Iddekinge, C.H., Putka, D.J., & Sager, C.E. (2005). Attitudinal criteria. In D.J. Knapp & T.R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (pp. 89-104) (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

White, L.A., & Young, M.C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

White, L.A., Young, M.C., Hunter, A.E., & Rumsey, M.G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(3), 291-295.

APPENDIX A: TAPAS FORM EQUIVALENCE ANALYSES

PSYCHOMETRIC EVALUATION OF THE TAPAS[17]

Matthew T. Allen, Michael J. Ingerick, and Justin A. DeSimone (HumRRO)

The purpose of this appendix is to present a psychometric evaluation of the TAPAS in an applicant setting.[18] Specifically, we begin by comparing the psychometric characteristics (means, standard deviations, and intercorrelations) of the three versions of the TAPAS to one another. This is followed by an empirical comparison of the TOPS versions of the TAPAS with the TAPAS-95s, which was administered as part of the EEEM research (see Chapter 1).

Empirical Comparison of the Three TAPAS Versions

As described in Chapter 3, three versions of the TAPAS were administered as part of the TOPS research: (a) a computer-adaptive 13-dimension version (13D-CAT), (b) a static 15-dimension version (15D-Static), and (c) a computer-adaptive 15-dimension version (15D-CAT). Although the three versions were intended to be comparable, they should not be seen as parallel. All versions were based on the same statement pool, but the dimensionality, test length, and/or design specifications (i.e., the blueprints) varied. To determine whether the three versions were sufficiently equivalent to treat as one measure in subsequent analyses, we compared the three versions based on the (a) mean dimension scores and standard deviations and (b) intercorrelations among the dimension scores. The means and standard deviations of the raw dimension scores for the three TAPAS versions are summarized in Table A.1. To compare the magnitude of the mean differences, standardized mean differences (i.e., Cohen's *d*) were computed for each TAPAS scale using the following formula:

$$d = M_{GROUP1} - M_{GROUP2} / SD_{POOLED} \quad (1)$$

Cohen's (1988) rule of thumb suggests that 0.20 to 0.30 should be considered a small effect, 0.50 a medium effect, and 0.80 or above a large effect. The differences between standard deviations were compared with an *F*-test, which was computed with the following formula:

$$F = SD^2_{GROUP1} / SD^2_{GROUP2} \quad (2)$$

We did not compute statistical significance tests for either the mean or standard deviation differences due to the large sample sizes of the three groups. Because of the large sample sizes, even small differences would be considered statistically significant using traditional null hypothesis testing. Accordingly, we focused on the effect size estimates when comparing the three versions.

[17] This text was taken nearly verbatim from Allen et al., 2011. It has been lightly edited to fit the context of the current report.

[18] Although operational in the strictest sense of the term, the TOPS IOT&E applies a low screen to such few applicants (those in Education Tier 1 scoring in AFQT Category IV), that applicants may recognize that the scores are unlikely to matter for them personally. Thus, we use the term "applicant" rather than "operational" setting.

Table A.1. Standardized Mean Score and Standard Deviation Differences between TOPS IOT&E TAPAS Versions by Scale

	TOPS TAPAS Version											
	13D-CAT (n = 1,311)		15D-Static (n = 8,224)		15D-CAT (n = 42,130)		Cohen's d			F-test		
Composite/Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	$d_{13D\text{-}15DS}$	$d_{13D\text{-}15C}$	$d_{15DS\text{-}15C}$	$F_{15DS\text{-}13D}$	$F_{13D\text{-}15C}$	$F_{15DS\text{-}15C}$
By Individual Composite/Scale												
Achievement	.234	0.493	.275	0.503	.150	0.480	-0.08	0.18	0.26	0.96	1.05	1.09
Adjustment	--	--	.159	0.582	-.005	0.570	--	--	0.29	--	--	1.04
Attention Seeking	-.224	0.557	-.246	0.528	-.194	0.533	0.04	-0.06	-0.10	1.11	1.09	0.98
Cooperation	.027	0.390	-.070	0.392	-.061	0.375	0.25	0.23	-0.03	0.99	1.08	1.10
Dominance	.072	0.600	-.026	0.589	.035	0.591	0.17	0.06	-0.10	1.04	1.03	1.00
Even Tempered	.126	0.514	.253	0.480	.159	0.477	-0.26	-0.07	0.20	1.15	1.16	1.01
Generosity	-.172	0.426	-.196	0.449	-.203	0.430	0.05	0.07	0.02	0.90	0.98	1.09
Intellectual Efficiency	.099	0.608	-.086	0.593	-.018	0.587	0.31	0.20	-0.12	1.05	1.07	1.02
Non-Delinquency	.105	0.457	.117	0.457	.088	0.459	-0.03	0.04	0.06	1.00	0.99	0.99
Optimism	.175	0.464	.261	0.511	.134	0.462	-0.17	0.09	0.27	0.83	1.01	1.22
Order	-.416	0.568	-.397	0.575	-.431	0.548	-0.03	0.03	0.06	0.98	1.08	1.10
Physical Conditioning	-.019	0.617	-.048	0.619	.026	0.629	0.05	-0.07	-0.12	1.00	0.96	0.97
Self Control	--	--	.098	0.527	.058	0.532	--	--	0.07	--	--	0.98
Sociability	-.026	0.622	-.209	0.594	-.037	0.594	0.31	0.02	-0.29	1.09	1.09	1.00
Tolerance	-.240	0.598	-.249	0.588	-.231	0.570	0.02	-0.02	-0.03	1.03	1.10	1.06
Can-Do Composite	.739	1.406	.821	1.382	.513	1.373	-0.06	0.16	0.22	1.03	1.05	1.01
Will-Do Composite	.669	1.319	.844	1.225	.616	1.247	-0.14	0.04	0.18	1.16	1.12	0.96
Averages	*M*	*SD*	*M*	*SD*	*M*	*SD*	$\lvert d \rvert$	$\lvert d \rvert$	$\lvert d \rvert$	*F*	*F*	*F*
All TAPAS Scales	-.020	0.532	-.024	0.532	-.035	0.522	0.12	0.08	0.13	1.01	1.05	1.04

Note. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above. 13D = TAPAS 13D-CAT, 15DS = TAPAS 15D-Static, 15C = TAPAS 15D-CAT.

The results in Table A.1 suggest that despite the differences in length, dimensionality, and design specifications acknowledged at the outset, the three versions of the TAPAS were quite similar in terms of their means and standard deviations. The *d* statistics ranged from a low of 0.02 to a high of 0.31 and the average absolute values of the *d* statistics were all below 0.20, which is considered a "small" difference. Only 7 of the 47 pairwise comparisons were above 0.25, or a quarter of a standard deviation. Three of these differences were between the 13D-CAT and 15D-Static, and four were between the 15D-CAT and 15D-Static. The 13D-CAT and 15D-CAT versions had the most similar means. Overall, this suggests that the number of dimensions (13 or 15) and format (static or adaptive) of the TAPAS had little effect on the facet mean and standard deviation scores, though the format led to slightly more differences. The largest differences tended to be for the Sociability, Intellectual Efficiency, Optimism, and Cooperation scales. In terms of standard deviations, all of the *F*-values were near 1.0, suggesting that the variances are roughly equivalent across the three versions. The one exception to this pattern was the Optimism scale, which exhibited an *F* value of 1.22 between the 15D-Static and 15D-CAT versions of the TAPAS.

Another basis for examining the consistency between the different TAPAS versions is in the pattern of intercorrelations among the dimension scores. For example, if Dominance is positively correlated with Achievement in one version of the TAPAS, we would reasonably expect a positive correlation of a similar magnitude to be found in another version of the TAPAS, regardless of any mean score differences between the versions. Specifically, we would expect a similar pattern of intercorrelations among the dimensions that are theoretically or taxonomically related, such as the facets underlying the Big Five (see Table 3.1, Chapter 3). To test the similarity of the intercorrelation matrices for the three versions, we computed a Standardized Root Mean Square Residual (SRMR). Following Hu and Bentler (1999), the SRMR was computed using the following formula,

$$\text{SRMR} = \sqrt{\left\{ 2 \sum_{i=1}^{p} \sum_{j=1}^{i} [(s_{ij} - \hat{\sigma}_{ij}) / (s_{ii}\, s_{jj})]^2 \right\} / p(p+1)} \qquad (3)$$

where s_{ij} is the observed covariances for one group (i.e., applicants completing one TAPAS version), σ_{ij} is the observed covariances for the comparison group, s_{ii} and s_{jj} are the observed standard deviations, and p is the number of observed variables. SRMR is a commonly used fit index in confirmatory factor analysis. Following Hu and Bentler's (1999) recommendations, we interpreted SRMRs that are close to zero as very similar, while those above .08 were interpreted as different.

The results of the SRMR analysis can be found in Table A.2. We report the SRMRs comparing (a) the full correlation matrices, (b) the matrices corresponding to the Big Five, and (c) the matrices corresponding to the can-do and will-do TAPAS composites. The SRMRs based on (b) and (c) were computed to better diagnose where the systematic differences, if any, were among the versions that may otherwise be lost from an examination of the full matrices. Overall, the results suggest that the patterns of intercorrelations were very similar between the three TAPAS versions. No SRMR values were above .08, and only one SRMR value – comparing the 13D-CAT and 15D-Static versions – was above .05. Further examination of the bivariate correlations between the two versions suggests that the main sources of discrepancy were on the Achievement, Cooperation, and Even Tempered scales. For example, the Achievement/Cooperation ($r_{13D\text{-}CAT}$ = .07, $r_{15D\text{-}Static}$ = -.03; *Z*

= 3.26, $p < .01$) and Achievement/Optimism ($r_{13D\text{-}CAT} = .13$, $r_{15D\text{-}Static} = .26$; $Z = -4.33$, $p < .01$) correlations were significantly different between the two versions. Overall, however, the results of the SRMR analysis suggest the patterns of intercorrelations for the two versions are quite similar.

Table A.2. Standardized Differences in Scale Score Intercorrelations between the TOPS IOT&E TAPAS Versions by Dimension

Composite/ Scale Score Profile	$SRMR_{13C\text{-}15DS}$	$SRMR_{13C\text{-}15C}$	$SRMR_{15DS\text{-}15C}$
All TAPAS Scales	.0574	.0357	.0468
By Big Five Factor			
Agreeableness	.0059	.0019	.0040
Conscientiousness	.0243	.0243	.0246
Emotional Stability	.0060	.0178	.0370
Extraversion	.0348	.0259	.0182
Openness to Experience	.0169	.0061	.0107
By TOPS Composite			
Can-Do	.0480	.0208	.0395
Will-Do	.0475	.0280	.0276

Note. 13D-CAT, n = 1,311. 15D-Static, n = 8,224. 15D-CAT, n = 42,130. Values reported are standardized root mean squared residuals (SRMR). SRMR values greater than .08 are bolded (Hu & Bentler, 1999). Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above. 13D = TAPAS 13D-CAT, 15DS = TAPAS 15D-Static, 15C = TAPAS 15D-CAT.

In summary, the results suggest that the means, standard deviations, and intercorrelations for the three versions are comparable. Therefore, it would be appropriate to combine scores from the three versions in the same analysis, provided that the scores are standardized within version to account for any small scaling differences.

Comparison of the TAPAS-95s with the TOPS IOT&E TAPAS

Previous work testing temperament measures such as the Assessment of Individual Motivation (AIM) under operational conditions has found high levels of socially desirable responding that lead to criterion-related validity coefficients approaching zero[19] (White, Young, Hunter, & Rumsey, 2008). These results motivated the continuation of research on fake-resistant personality measures and, in fact, led to the development of the TAPAS. For obvious reasons, the critical evidence concerning the effectiveness of TAPAS for selection applications lies in its performance under operational conditions, so comparisons of internal and relational properties across examinee groups taking the test in a common environment (e.g., military entrance processing stations) but under different instructions (operational vs. research only) would be highly valued. Because the data for such comparisons were not available for this report, an alternative was to compare means, intercorrelations, and validities for the three versions of TAPAS explored in the IOT&E to the TAPAS-95s administered in the EEEM research project (Knapp & Heffner, 2010). Despite the systematic differences in the examinee pools and test forms discussed previously, such analyses were seen as useful for providing at least a rough indication of the effect of situational factors on the test scores.

[19] Additional work with the AIM as a component of the Tier Two Accession Screen (TTAS) has demonstrated that it contributes to the prediction of attrition for applicants who are non-high school diploma graduates.

To address this issue, we conducted analyses similar to those from the previous section. Specifically, we compared TOPS TAPAS versions to the TAPAS-95s based on (a) the facet score means and standard deviations, (b) intercorrelations among facet scores, and (c) correlations between the dimension scores and external individual difference variables (e.g., demographics, AFQT scores). To ensure that the two samples were as comparable as possible, the results of the TOPS TAPAS analyses were limited to respondents that were Education Tier 1, non-prior service, AFQT Category IV and above, and had signed a contract with the Army. The results for the TAPAS-95s were also limited to Education Tier 1, non-prior service Soldiers. Since the TAPAS-95s was administered to new accessions, the sample was already restricted to individuals who had signed a contract and were AFQT Category IV or above.

It is important to note that the TOPS TAPAS versions and TAPAS-95s are not parallel measures because many statements used in the TAPAS-95s were also included in the TOPS TAPAS statement *pool*, but parameters for some statements were re-estimated in accordance with refinements to the TAPAS trait taxonomy. For example, statements from the TAPAS-95s "Optimism" facet were reallocated to the "Adjustment" and "Optimism" facets before the TOPS implementation. In addition, statement parameters for Tolerance, Order, Cooperation, and Even Tempered were revised based on additional data that were collected, thus making direct comparisons between TOPS and EEEM difficult.

In sum, substantive differences between the EEEM context and the present one are enumerated below.

1. The TAPAS-95s was administered via paper and pencil, while the three versions of the TOPS TAPAS were computer-administered.

2. The TAPAS-95s was static, while two of the three TOPS TAPAS versions were adaptive.

3. The TAPAS-95s assessed 12 dimensions using 95 items, whereas the TOPS TAPAS versions assessed 13 dimensions with 104 items or 15 dimensions with 120 items.

4. The TAPAS-95s was administered to Soldiers who had already accessed into the Army, whereas the TOPS TAPAS versions were administered to an applicant sample.

5. The TAPAS-95s was administered in an environment where the Army was having difficulty meeting its recruiting mission, whereas TOPS TAPAS was administered in a poor economic environment (McMichael, 2008; 2009; Schafer, 2007) in which recruiting was less challenging. As a result of these economic conditions, the Army became more selective in its recruiting and accessioning process during the course of the TOPS research.

Despite these aforementioned differences, substantial score inflations in operational settings and/or large changes in intercorrelations or correlations with external variables for the TOPS TAPAS versions could signal that the test is functioning differently as compared to research settings.

Table A.3 presents the mean and standard deviations for 10 scales found in both TAPAS-95s and the three TOPS TAPAS versions. The scales with the smallest standardized mean differences were the Achievement (Avg. $|d| = 0.12$), Attention Seeking (Avg. $|d| = 0.13$), Non-Delinquency (Avg. $|d| = 0.02$), and Physical Conditioning (Avg. $|d| = 0.20$) scales. The Tolerance (Avg. $|d| = 0.27$) and Dominance (Avg. $|d| = 0.27$) scales also had standardized mean differences below 0.30. The Even Tempered (Avg. $|d| = 1.12$) and Order (Avg. $|d| = 0.70$) scales evidenced the largest mean differences, but these scores were based on parameters that were updated prior to TOPS, so the difference in means might be explained to some extent by changes in the IRT metrics. Also, certain facet scores such as Physical Conditioning decreased for the TOPS TAPAS as compared to TAPAS-95s, which would not be expected if faking were present.

The standard deviations of the TOPS TAPAS dimension scores reported in Table 4.3 were generally lower than the corresponding standard deviations observed on the TAPAS-95s. The average *F* values reflecting the difference in the standard deviations between the three TOPS TAPAS versions and the TAPAS-95s were consistently close to 2.00. With regard to scores on the individual dimensions, the Tolerance (Avg. $F = 1.31$), Intellectual Efficiency (Avg. $F = 1.21$), and Dominance (Avg. $F = 1.01$) standard deviations were most similar between the two settings, while the Cooperation (Avg. $F = 5.17$), Attention Seeking (Avg. $F = 2.22$), Even Tempered (Avg. $F = 2.51$), and Non-Delinquency (Avg. $F = 2.22$) scores demonstrated the largest differences. The magnitude and pattern of the differences in the standard deviations between the two settings were generally the same across the three TOPS TAPAS versions.

Another way to compare TAPAS-95s and TOPS TAPAS is to examine the consistency of their relationship with each other and with key individual difference variables. Correlations are useful because they are unaffected by linear transformations, associated with, for example, changing means or IRT recalibrations. Marked differences across settings or versions of a test could provide insights into how test construction practices affect item responding and ultimately construct and predictive validities.

With this in mind, we compared the patterns of intercorrelations among the facet scores from the TAPAS-95s to those observed in the TOPS TAPAS using the SRMR statistic described previously. The SRMR results are reported in Table A.4. Note that we did not compute SRMRs for the Agreeableness and Emotional Stability dimensions because there was only one scale in each that was included in both the TOPS and EEEM studies. The Optimism scale was excluded from these analyses due to the content changes described above. Overall, we found few differences between the three TOPS TAPAS versions and the TAPAS-95s within the groupings where we would expect the most stable relationships (i.e., within Big Five dimension). The differences in matrices for the can-do composite were also relatively small. The larger differences in the two matrices were found for all of the TAPAS scales and the will-do composite.

Table A.3. Standardized Mean Score and Standard Deviation Differences between EEEM TAPAS-95s and the TOPS IOT&E TAPAS by Version and Scale

	TAPAS Version													
	EEEM (95s) (n = 3,381)		13D-CAT (n = 786)				15D-Static (n = 4,258)				15D-CAT (n = 18,217)			
Scale	*M*	*SD*	*M*	*SD*	*d*	*F*	*M*	*SD*	*d*	*F*	*M*	*SD*	*d*	*F*
Achievement	.160	.625	.230	.495	0.12	1.59	.286	.512	0.22	1.49	.157	.481	-0.01	1.69
Attention Seeking	-.127	.797	-.206	.554	-0.10	2.07	-.236	.521	-0.17	2.35	-.192	.534	-0.11	2.23
Cooperation	-.282	.865	.027	.375	0.39	5.32	-.089	.390	0.30	4.92	-.048	.377	0.48	5.26
Dominance	-.144	.603	.070	.591	0.36	1.04	-.045	.608	0.16	0.98	.028	.600	0.29	1.01
Even Tempered	-.491	.764	.145	.497	0.88	2.36	.261	.479	1.21	2.55	.181	.473	1.27	2.61
Intellectual Efficiency	-.187	.647	.121	.596	0.48	1.18	-.046	.589	0.23	1.20	.011	.579	0.34	1.25
Non-Delinquency	.120	.661	.128	.430	0.01	2.37	.128	.448	0.01	2.18	.107	.455	-0.03	2.11
Order	-.034	.636	-.464	.560	-0.69	1.29	-.427	.574	-0.65	1.23	-.462	.551	-0.76	1.33
Physical Conditioning	.128	.712	.000	.609	-0.19	1.37	-.040	.627	-0.25	1.29	.033	.626	-0.15	1.29
Tolerance	-.420	.673	-.261	.599	0.24	1.26	-.259	.591	0.26	1.30	-.238	.575	0.31	1.37

Note. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above, signed contract.

The main source of difference in the intercorrelation matrices most often involved the Attention Seeking (Avg. Total Δ| *r* | = .10) scale, which had the largest average difference between the three versions of the TOPS TAPAS and the TAPAS-95s. For example, the Attention Seeking scale had four intercorrelations where the average difference was above .10: (a) Cooperation (Avg. Δ| *r* | = .11), (b) Intellectual Efficiency (Avg. Δ| *r* | = .13), (c) Non-Delinquency (Avg. Δ| *r* | = .23), and (d) Achievement (Avg. Δ| *r* | = .15). Other scales that had large average differences include Cooperation (Avg. Total Δ| *r* | = .09), and Dominance (Avg. Total Δ| *r* | = .08). The Order (Avg. Total Δ| *r* | = .03) and Physical Conditioning (Avg. Total Δ| *r* | = .05) scales had the smallest average differences.

Table A.4. Standardized Differences in Scale Score Intercorrelations between the EEEM TAPAS-95s and the TOPS IOT&E TAPAS by Version and Dimension

	TAPAS Version		
Composite/ Scale Score Profile	13D-CAT (*n* = 786)	15D-Static (*n* = 4,258)	15D-CAT (*n* = 18,217)
All TAPAS Scales	.0754	**.0800**	**.0810**
By Big Five Factor			
Agreeableness	n/a	n/a	n/a
Conscientiousness	.0166	.0151	.0192
Emotional Stability	n/a	n/a	n/a
Extraversion	.0305	.0687	.0339
Openness to Experience	.0442	.0344	.0420
By TOPS Composite			
Can-Do	.0471	.0630	.0450
Will-Do	.0600	**.0984**	**.0867**

Note. Values reported are standardized root mean squared residuals (SRMR). SRMR values greater than .08 are bolded (Hu & Bentler, 1999). Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above, signed contract. The raw TAPAS scores were used in this analysis.

We also computed the correlations (or point-biserial correlation for binary demographic variables) between TAPAS dimension scores and four variables: (a) AFQT score, (b) race, (c) ethnicity, and (d) gender. For this analysis, the TOPS TAPAS versions were combined into one overall set of TAPAS scales by:

1. Filtering out participants that were not part of the sample of interest (i.e., those that were not in Tier 1, non-prior service, AFQT Category IV or above; and

2. Standardizing the variables within version using a *z*-transformation, completed by subtracting each score from the mean for that version and dividing by the standard deviation.

Once the correlations were computed, the TAPAS-95s and TOPS TAPAS results were compared using two statistics. The first was the squared difference between the correlations (Δr^2). The

second was Fisher's *Z* test of the equality of two correlations, which can be expressed with the following formula (Cohen, Cohen, Aiken, & West, 2003):[20]

$$Z = \frac{z'_1 - z'_2}{\sqrt{1/(n_1-3) + 1/(n_2-3)}} \quad (4)$$

where z'_1 and z'_2 are the logarithmic transformations of the correlations for groups 1 and 2 and n_1 and n_2 are the sample sizes. Values above 1.96 or less than -1.96 are considered statistically significant.

The results of this analysis are presented in Table A.5. We generally found weak relations between TAPAS dimension scores and key individual difference variables. Very few of the correlations were above .10, and many were not statistically significant despite the large sample sizes. However, there were exceptions. For example, Intellectual Efficiency and the can-do TAPAS composite (which includes the Intellectual Efficiency scale) were strongly related to Soldiers' AFQT scores in both the EEEM TAPAS-95s and TOPS TAPAS versions. Second, Tolerance was positively correlated with all three demographic variables (race, ethnicity, and gender), suggesting that minority subgroups (Blacks, Hispanics, and females) tended to score higher on the Tolerance scale than the majority subgroups (Whites, Non-Hispanics, and males). The Generosity scale was also positively correlated with gender, suggesting that females score higher on that scale than males, while Physical Conditioning was negatively correlated with gender, suggesting that males score higher on that scale than females. While there were a number of other statistically significant correlations between the individual correlations and these demographic variables, the magnitude was generally small. This finding is further supported by the subgroup mean differences, presented as a reference in Appendix B.

There were differences between the TAPAS-95s and the TOPS TAPAS, as measured by the Δr^2 estimates, but the Δr^2 values were consistently .03 or less. Although a number of the *Z* comparisons were statistically significant, this was likely primarily due to the large sample sizes available for these analyses. The correlations demonstrating the largest differences between the two settings involved the Attention Seeking scale with AFQT, and the Dominance scale with gender. The Attention Seeking scale was negatively correlated with AFQT in EEEM, and positively correlated with AFQT in TOPS. Finally, the Dominance scale was positively correlated with gender in EEEM, and negatively correlated in TOPS. Despite these apparent differences, there was no systematic pattern of results to suggest that the relationship between these individual difference variables and the TAPAS changed fundamentally from one setting to the other.

[20] Note that Fisher's *Z* assumes that the variables under consideration are normally distributed. However, the dichotomous variables used in this analysis (race, ethnicity, and gender) are not normally distributed and, therefore, violate this assumption. Nevertheless, given that the purpose of this analysis was to measure the relative magnitude of the difference between two coefficients and that the Fisher's *Z* is appropriate for the AFQT/TAPAS correlations, the Fisher's *Z* was used for the dichotomous variables as well. However, this limitation should be kept in mind when interpreting these results.

Table A.5. Differences in Scale Score Correlations between the TAPAS-95s and the TOPS IOT&E TAPAS with Individual Difference Variables

	EEEM				TOPS (Standardized)				Difference Metrics							
Scale	AFQT *r*	Race *r*	Eth *r*	Sex *r*	AFQT *r*	Race *r*	Eth *r*	Sex *r*	AFQT $(\Delta r)^2$	Race $(\Delta r)^2$	Eth $(\Delta r)^2$	Sex $(\Delta r)^2$	AFQT *Z*	Race *Z*	Eth *Z*	Sex *Z*
Achievement	**.06**	-.02	-.01	**.08**	**.07**	**-.02**	**-.02**	.00	.00	.00	.00	.01	-0.71	-0.28	0.70	**4.17**
Adjustment	.	.	.	.	**.09**	**-.04**	**-.06**	**-.11**	.	.	.	.	.	.	.	.
Attention Seeking	**-.07**	-.02	-.01	-.01	**.09**	**-.03**	**-.02**	**-.04**	.03	.00	.00	.00	**-8.97**	0.52	0.32	1.35
Cooperation	**-.04**	-.01	.00	**.05**	-.01	.01	.00	.00	.00	.00	.00	.00	-1.39	-1.05	0.00	**2.36**
Dominance	**.06**	**.08**	-.01	**.09**	**.07**	.00	**.02**	**-.05**	.00	.01	.00	.02	-0.90	**3.81**	-1.36	**7.48**
Even Tempered	**.14**	.02	.00	**-.04**	**.06**	.01	**-.02**	**-.03**	.01	.00	.00	.00	**3.96**	0.78	0.90	-0.63
Generosity	.	.	.	.	**-.07**	**.05**	**.02**	**.15**	.	.	.	.	.	.	.	.
Intellectual Efficiency	**.38**	.02	-.04	**-.07**	**.42**	**-.03**	**-.05**	**-.07**	.00	.00	.00	.00	-2.12	**2.91**	0.87	0.38
Non-Delinquency	**.06**	-.01	-.04	**.14**	.00	.01	**-.03**	**.05**	.00	.00	.00	.01	**2.83**	-1.27	-0.35	**4.57**
Optimism	.	.	.	.	.00	.00	.00	-.01	.	.	.	.	.	.	.	.
Order	**-.04**	**.06**	.02	**.11**	**-.15**	**.07**	**.05**	**.05**	.01	.00	.00	.00	**6.27**	-0.45	-1.37	**3.29**
Physical Conditioning	.00	.01	.01	**-.12**	**.03**	**-.06**	**-.03**	**-.14**	.00	.00	.00	.00	-1.52	**3.39**	1.79	1.18
Self Control	.	.	.	.	.00	**.07**	**.03**	.01	.	.	.	.	.	.	.	.
Sociability	.	.	.	.	**-.09**	**-.02**	.00	.01	.	.	.	.	.	.	.	.
Tolerance	.02	**.12**	**.08**	**.10**	**-.01**	**.08**	**.10**	**.13**	.00	.00	.00	.00	1.80	1.84	-1.21	-1.67

Note. EEEM AFQT *n* = 3,362, EEEM Race *n* = 3,194, EEEM Ethnicity *n* = 2,833, EEEM Gender *n* = 3,368. TOPS AFQT *n* = 22,475-23,261, TOPS Race *n* = 16,909-17,416, TOPS Ethnicity *n* = 18,166-18,649, TOPS Gender *n* = 22,475-23,261. All of the demographic variables were coded as 1 or 0, with 1 being the minority subgroup: Race (1=Black, 0=White), Ethnicity (1=Hispanic, 0=Non-Hispanic), and Gender (1=Female, 0=Male). Δr^2 = the squared difference between the TOPS and EEEM TAPAS correlations. *Z* = The difference between the TOPS and EEEM TAPAS correlations as determined using Fisher's *Z* test. Values above 1.96 are bolded. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above, signed contract.

Summary

To test whether the psychometric characteristics of the TAPAS were consistent across versions (13D-CAT, 15D-Static, 15D-CAT) and settings (EEEM vs. IOT&E), we conducted a number of diagnostic and comparative analyses. The results of these analyses suggest:

1. The three versions of the TAPAS (13D-CAT, 15D-Static, and 15D-CAT) were consistent with one another in terms of their means, standard deviations, and patterns of intercorrelations. The two computer-adaptive versions of the TAPAS were particularly similar. However, there were some mean differences for individual scales, suggesting the need to standardize within these three versions to account for scaling differences if versions other than the 15D-CAT are used in future assessments.

2. The standard deviations for the TOPS TAPAS were, on average, smaller than in the EEEM research, suggesting either (a) the TOPS population is narrower on these facets or (b) participants are responding in a way that is reducing the available variance for each scale.

3. Some of the TAPAS scales were more similar across the research and operational settings than others. For example, the psychometric properties for the Attention Seeking scale changed substantially from one setting to another, while the Tolerance and Physical Conditioning scales were similar across the two settings.

4. With a few exceptions, the TAPAS scales showed no bias as they were not strongly related to key individual difference variables (AFQT scores, race, ethnicity, and gender). Additionally, the patterns of these relationships were generally consistent from the EEEM to TOPS settings.

Keeping in mind that previous research has shown large differences between the experimental and operational use of temperament measures (White et al., 2008), these results suggest that the use of the TAPAS in an operational setting is promising. Although there were some differences in scale score means and standard deviations across the two settings, these differences could be explained by differences in test specifications and IRT metrics or other environmental factors rather than socially desirable responding.

APPENDIX B: PREDICTOR DESCRIPTIVE STATISTICS

Table B.1. Raw Mean and Standard Deviations for the TOPS IOT&E TAPAS Scales by Version

	TOPS TAPAS Version					
	13D-CAT (n = 1,312)		15D-Static (n = 9,967)		15D-CAT (n = 73,212)	
TAPAS Composite/Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*
By Individual Composite/Scale						
Achievement	.234	.493	.276	.500	.154	.482
Adjustment	.	.	.160	.586	-.011	.570
Attention Seeking	-.223	.557	-.252	.529	-.196	.531
Cooperation	.027	.390	-.071	.391	-.065	.372
Dominance	.072	.600	-.023	.589	.034	.587
Even Tempered	.126	.515	.253	.482	.155	.477
Generosity	-.171	.426	-.194	.447	-.204	.430
Intellectual Efficiency	.100	.608	-.087	.592	-.016	.586
Non-Delinquency	.104	.457	.121	.453	.088	.460
Optimism	.175	.464	.269	.507	.133	.461
Order	-.415	.569	-.398	.572	-.424	.546
Physical Conditioning	-.018	.616	-.044	.618	.034	.627
Self Control	.	.	.094	.526	.061	.532
Sociability	-.025	.621	-.210	.596	-.040	.593
Tolerance	-.239	.598	-.257	.589	-.229	.569
Can-Do Composite	-.007	2.707	.037	2.666	.003	2.744
Will-Do Composite	-.002	2.473	.029	2.347	.018	2.399

Note. Results are limited to the Applicant Sample (Non-prior service, Education Tier 1, AFQT Category IV and above). 13D = TAPAS 13D-CAT, 15DS = TAPAS 15D-Static, 15C = TAPAS 15D-CAT.

Table B.2. Predictor Intercorrelations

TAPAS Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. AFQT																	
2. Achievement	**.10**																
3. Adjustment	**.11**	**.09**															
4. Attention Seeking	**.11**	**.05**	**.11**														
5. Cooperation	.00	**.09**	**.12**	**.05**													
6. Dominance	**.09**	**.33**	**.11**	**.21**	-.01												
7. Even Tempered	**.08**	**.10**	**.19**	**-.01**	**.24**	**-.06**											
8. Generosity	**-.07**	**.08**	**-.02**	**-.08**	**.19**	.00	**.11**										
9. Intellectual Efficiency	**.43**	**.26**	**.18**	**.08**	**.02**	**.25**	**.08**	**-.02**									
10. Non-Delinquency	-.01	**.18**	.00	**-.13**	**.16**	**-.01**	**.18**	**.14**	**.01**								
11. Optimism	**.01**	**.20**	**.28**	**.17**	**.16**	**.17**	**.18**	**.04**	**.10**	**.08**							
12. Order	**-.18**	**.16**	**-.08**	**-.09**	.00	**.05**	**-.03**	**.04**	**.01**	**.09**	**-.01**						
13. Physical Condition	**.05**	**.15**	**.07**	**.12**	**-.02**	**.18**	**-.08**	**-.04**	**.05**	**-.03**	**.10**	**.03**					
14. Self Control	**-.01**	**.21**	**.06**	**-.12**	**.12**	**.03**	**.19**	**.08**	**.16**	**.24**	**.06**	**.18**	**-.06**				
15. Sociability	**-.08**	**.05**	**.12**	**.36**	**.18**	**.22**	**.03**	**.06**	.00	**-.05**	**.23**	**-.04**	**.13**	**-.12**			
16. Tolerance	**-.01**	**.11**	**.01**	**.02**	**.14**	**.06**	**.12**	**.32**	**.07**	**.07**	**.08**	**.03**	**-.06**	**.11**	**.11**		
17. Can-Do	**.22**	**.62**	**.27**	**.06**	**.24**	**.25**	**.55**	**.13**	**.52**	**.52**	**.56**	**.08**	**.07**	**.31**	**.10**	**.16**	
18. Will-Do	**.05**	**.56**	**.10**	**-.39**	**.18**	**.09**	**.49**	**.15**	**.13**	**.59**	**.16**	**.14**	**.38**	**.29**	**-.08**	**.09**	**.70**

Note. N = 83,179-88,017. Coefficients in bold are statistically significant, $p < .05$. Results are limited to the Applicant Sample (Non-prior service, Education Tier 1, AFQT Category IV and above).

Table B.3. TOPS Subgroup Mean Differences for Applicant Sample

	Ethnicity					Race					Gender				
	Non-Hispanic (NH)		Hispanic (H)		NH-H	White (W)		Black (B)		W-B	Male (M)		Female (F)		M-F
Scale/Predictor	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Standardized TAPAS Scales															
Achievement	0.01	1.01	-0.06	0.95	**0.07**	0.02	1.01	-0.06	0.95	**0.08**	0.00	1.01	0.02	0.96	-0.02
Adjustment	0.02	1.01	-0.12	0.95	**0.14**	0.03	1.01	-0.07	0.98	**0.10**	0.06	1.00	-0.23	0.97	**0.30**
Attention Seeking	0.01	1.01	-0.03	0.96	**0.04**	0.02	1.01	-0.08	0.93	**0.09**	0.02	1.00	-0.09	0.99	**0.12**
Cooperation	0.01	1.00	-0.03	0.97	**0.04**	-0.01	1.00	0.01	0.98	-0.01	0.00	1.00	-0.01	1.00	0.01
Dominance	0.00	1.01	0.01	0.94	-0.02	0.01	1.02	0.03	0.90	**-0.02**	0.03	1.01	-0.10	0.97	**0.13**
Even Tempered	0.02	1.01	-0.08	0.95	**0.10**	0.01	1.00	0.00	0.99	0.01	0.02	1.00	-0.08	1.01	**0.10**
Generosity	-0.01	1.01	0.05	0.96	**-0.06**	-0.02	1.01	0.11	0.97	**-0.14**	-0.08	0.99	0.32	0.98	**-0.40**
Intellectual Eff.	0.02	1.01	-0.13	0.93	**0.15**	0.03	1.01	-0.11	0.92	**0.14**	0.04	1.01	-0.16	0.93	**0.20**
Non-Delinquency	0.01	1.00	-0.04	0.97	**0.05**	0.01	1.00	0.05	0.97	**-0.04**	-0.03	1.01	0.12	0.97	**-0.15**
Optimism	0.00	1.01	0.01	0.95	0.00	0.01	1.00	0.02	0.96	-0.01	0.01	1.00	-0.04	1.01	**0.05**
Order	-0.02	1.01	0.13	0.96	**-0.15**	-0.05	1.00	0.19	0.96	**-0.24**	-0.03	0.99	0.12	1.04	**-0.16**
Physical Condition	0.02	1.01	-0.07	0.94	**0.09**	0.04	1.01	-0.17	0.95	**0.21**	0.08	0.99	-0.32	0.97	**0.41**
Self Control	-0.02	1.00	0.09	0.98	**-0.10**	-0.03	1.00	0.18	0.99	**-0.21**	-0.01	1.00	0.02	1.01	**-0.03**
Sociability	0.00	1.01	0.00	0.95	0.01	0.01	1.01	-0.05	0.94	**0.07**	0.00	1.00	0.00	0.99	0.00
Tolerance	-0.04	1.01	0.23	0.91	**-0.27**	-0.05	1.01	0.19	0.92	**-0.24**	-0.07	1.00	0.28	0.96	**-0.35**
Can-Do Composite	0.02	1.00	-0.11	0.97	**0.14**	0.03	1.00	-0.03	0.99	**0.06**	0.01	1.00	-0.05	1.01	**0.06**
Will-Do Composite	0.02	1.00	-0.10	0.97	**0.12**	0.02	1.00	-0.04	0.98	**0.07**	0.02	1.00	-0.07	1.01	**0.09**

Note. Ethnicity NH n = 64,393-65,427, H n = 12,145-12,188. Race W n = 60,595-61,492, B n = 11,693-11,854. Gender M n = 66,655-67,690, F n = 16,524-16,801. d = Standardized mean difference (Cohen's d). Results are limited to the Applicant sample (Non-prior service, Education Tier 1, AFQT Category IV and above). Coefficients in bold were statistically significant using an independent samples t-test ($p < .05$).

Table B.4. Descriptive Statistics for the ASVAB Based on the TOPS Applicant Sample

Measure/Scale	*n*	*M*	*SD*	*Min*	*Max*
AFQT	88,017	57.51	23.67	10	99
ASVAB Subtests					
General Science (GS)	84,791	51.77	8.58	20	76
Arithmetic Reasoning (AR)	84,791	52.69	7.87	22	72
Word Knowledge (WK)	84,791	51.44	8.30	18	76
Paragraph Comprehension (PC)	84,791	52.83	7.27	24	69
Math Knowledge (MK)	84,791	53.45	7.14	26	73
Electronics Information (EI)	84,791	52.26	9.23	16	84
Auto and Shop Information (AS)	84,791	50.44	9.55	20	86
Mechanical Comprehension (MC)	84,791	53.63	8.58	23	82
Assembling Objects (AO)	87,211	55.13	7.93	25	69
ASVAB Aptitude Area Composites					
Clerical (CL)	87,492	106.08	14.44	50	152
Combat (CO)	87,492	106.08	15.34	53	160
Electronics (EL)	87,492	105.86	15.35	52	160
Field Artillery (FA)	87,492	106.23	15.27	53	159
General Maintenance (GM)	87,492	105.64	15.78	53	161
Mechanical Maintenance (MM)	87,492	104.92	16.73	53	163
Operators and Food Service (OF)	87,492	105.63	15.76	52	160
Signal Communication (SC)	87,492	106.20	14.99	52	159
Skill Technical (ST)	87,492	106.04	15.01	51	157

Note. Results are limited to the Applicant Sample (non-prior service, Education Tier 1, AFQT Category IV and above).

APPENDIX C: DESCRIPTIVE STATISTICS FOR THE FULL SCHOOLHOUSE SAMPLE

Table C.1. Descriptive Statistics for Training Criteria Based on the Full Schoolhouse Sample

Measure/Scale	*n*	*M*	*SD*	*Min*	*Max*	*α*	*IRR*
Army Life Questionniare (ALQ)							
Affective Commitment[a]	17,572	3.86	0.68	1.00	5.00	.87	n/a
Normative Commitment[a]	17,572	4.17	0.70	1.00	5.00	.80	n/a
Career Intentions[b]	17,572	3.14	1.10	1.00	5.00	.92	n/a
Reenlistment Intentions[c]	17,572	3.61	0.98	1.00	5.33	.84	n/a
Attrition Cognition[d]	17,572	1.53	0.61	1.00	5.00	.78	n/a
Army Life Adjustment[a]	17,572	4.05	0.66	1.00	5.00	.85	n/a
Army Civilian Comparison[e]	17,572	3.92	0.71	1.00	5.00	.80	n/a
MOS Fit[a]	17,572	3.79	0.85	1.00	5.00	.92	n/a
Army Fit[a]	17,572	4.05	0.60	1.00	5.00	.86	n/a
Training Achievement[f]	17,550	0.39	0.59	0.00	2.00	n/a	n/a
Training Restarts[f]	17,572	0.42	0.65	0.00	4.00	n/a	n/a
Disciplinary Incidents[f]	7,009	0.24	0.57	0.00	6.00	n/a	n/a
Last APFT Score	17,322	248.19	32.59	5.00	300.00	n/a	n/a
MOS-Specific Job Knowledge Test (JKT)							
11B/11C/11X/18X	5,689	59.57	9.03	25.00	84.78	.76	n/a
31B	1,777	70.11	8.88	35.92	93.20	.79	n/a
68W	3,946	74.87	9.58	29.35	96.74	.85	n/a
88M	1,840	66.52	11.13	33.33	94.44	.78	n/a
91B	605	58.08	14.01	26.80	88.66	.91	n/a
WTBD Job Knowledge	16,990	65.65	12.46	10.00	100.00	.62	n/a
Army-Wide Performance Rating Scales[g]							
Effort	6,322	4.82	1.24	1.00	7.00	n/a	.19
Physical Fitness & Bearing	6,314	4.76	1.22	1.00	7.00	n/a	.19
Personal Discipline	6,376	4.90	1.27	1.00	7.00	n/a	.17
Commitment & Adjustment	6,356	4.96	1.22	1.00	7.00	n/a	.13
Support for Peers	6,305	4.95	1.18	1.00	7.00	n/a	.17
Peer Leadership	5,818	4.67	1.29	1.00	7.00	n/a	.19
Common Warrior Tasks Knowledge and Skill	5,851	4.82	1.20	1.00	7.00	n/a	.10
MOS Qualification Knowledge and Skills	5,545	4.89	1.16	1.00	7.00	n/a	.13
Overall Performance Scale	6,223	3.49	0.84	1.00	5.00	n/a	.27
MOS-Specific Performance Rating Composite Scores							
Total (combined across MOS)	4,012	4.55	0.94	1.00	7.00	n/a	n/a
11B/11C/11X/18X	1,849	4.80	0.87	1.00	7.00	.94	.14
31B	467	4.63	1.01	1.00	7.00	.95	.12
68W	1,910	4.35	0.84	1.00	7.00	.92	.04
88M	394	4.93	0.93	2.20	7.00	.92	.00
91B	127	4.78	1.78	1.00	7.00	.97	.11

Note. n/a = Internal consistency/coefficient alpha could not be computed for the scales/measures. Job knowledge scores are percent correct. WTBD = Warrior Tasks and Battle Drills. Results for 19K are not reported due to low sample size (n = 12). IRR = Interrater Reliability computed using G(q,k) (Putka, Le, McCloy, & Diaz, 2008).

[a] These items were responded to using agreement scales (1=Strongly Disagree to 5=Strongly Agree).

[b] This construct was measured by items using three types of scales: agreement scale (same as above), confident scale (1=Not At All Confident to 5=Extremely Confident), and likelihood scale (1=Extremely Unlikely to 5 = Extremely Likely)

[c] This construct was measured by items using agreement scale (same as above) and likelihood scale (same as above).

[d] This construct was measured by items using agreement scale (same as above) and often scale (1=Never to 5=Very Often).

[e] This construct was measured by the following scales: 1=Much Better in the Army, 2=Better in the Army, 3=About the Same, 4=Better in Civilian Life, 5=Much Better in Civilian Life.

[f] These scales are the total number of 'YES' responses to a series of yes/no questions about things that happened in training.

[g] The possible Army-wide and MOS-Specific Performance Rating Composite Scores are between 1 and 7.

Table C.2. Performance Rating Scales (PRS) Intercorrelations for Full Schoolhouse Sample

Scale	1	2	3	4	5	6	7	8	9
Army-Wide Performance Rating Scales (PRS)									
1. Effort									
2. Physical Fitness and Bearing	.69								
3. Personal Discipline	.73	.67							
4. Commitment and Adjustment	.71	.68	.78						
5. Support for Peers	.68	.62	.74	.77					
6. Peer Leadership	.65	.64	.66	.70	.72				
7. Common/Warrior Tasks Knowledge and Skills	.63	.63	.66	.72	.72	.73			
8. MOS Qualification Knowledge and Skills	.65	.64	.66	.73	.71	.67	.80		
9. Overall Performance	.58	.58	.57	.58	.53	.62	.54	.56	
MOS-Specific Performance Ratings Composites									
10. Combined MOS-Specific PRS	.53	.51	.53	.61	.58	.57	.68	.69	.48
11. 11B	.61	.57	.59	.66	.65	.63	.69	.71	.54
12. 31B	.66	.63	.69	.72	.70	.63	.74	.74	.64
13. 68W	.34	.29	.30	.39	.35	.47	.61	.50	.29
14. 88M	.54	.60	.61	.62	.64	.60	.68	.72	.54
15. 91B	.74	.76	.81	.85	.77	.72	.81	.89	.67

Note. All correlations are statistically significant ($p < .05$). Sample sizes for each research criterion variable can be found in Table 2.5.

Table C.3. Descriptive Statistics for Schoolhouse Criteria by MOS from the Full Schoolhouse Sample

	Total		11B		25U		31B		42A		68W		88M		91B	
Measure/Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Army Life Questionnaire (ALQ)																
Affective Commitment	3.86	0.68	3.91	0.66	3.58	0.72	3.96	0.63	3.88	0.67	3.74	0.71	3.95	0.65	3.81	0.69
Normative Commitment	4.17	0.70	4.20	0.68	3.97	0.80	4.20	0.68	4.07	0.75	4.15	0.71	4.22	0.69	4.03	0.75
Career Intentions	3.14	1.10	3.17	1.08	2.90	1.18	3.12	1.09	3.34	1.10	2.98	1.09	3.42	1.11	3.12	1.11
Reenlistment Intentions	3.44	0.98	3.46	0.94	3.17	1.05	3.41	0.97	3.54	0.96	3.33	0.98	3.67	0.96	3.39	0.98
Attrition Cognitions	1.53	0.61	1.51	0.62	1.74	0.72	1.50	0.57	1.58	0.63	1.56	0.60	1.49	0.60	1.64	0.68
Army Life Adjustment	4.05	0.66	4.05	0.68	3.99	0.62	4.09	0.62	4.02	0.63	4.01	0.64	4.10	0.63	4.02	0.70
Army Civilian Comparison	3.85	0.79	3.92	0.79	3.71	0.77	3.93	0.74	3.91	0.85	3.62	0.78	4.03	0.77	3.89	0.79
MOS Fit	3.79	0.85	3.87	0.80	3.34	0.84	3.91	0.78	3.55	0.91	3.97	0.82	3.32	0.85	3.62	0.91
Army Fit	4.05	0.60	4.08	0.59	3.84	0.63	4.14	0.55	4.10	0.61	3.94	0.61	4.16	0.58	3.97	0.62
Training Achievement	0.39	0.59	0.44	0.66	0.46	0.60	0.36	0.58	0.44	0.62	0.29	0.48	0.38	0.57	0.38	0.54
Training Restarts	0.42	0.65	0.30	0.55	0.72	0.81	0.26	0.51	0.71	0.79	0.56	0.74	0.47	0.67	0.57	0.72
Disciplinary Incidents	0.24	0.57	0.25	0.58	--	--	--	--	--	--	--	--	--	--	--	--
Last APFT Score	248.19	32.59	244.90	33.11	243.56	34.39	258.30	30.08	242.87	33.31	249.71	31.82	245.76	31.94	243.62	29.60
MOS-Specific JKT	--	--	59.57	9.03	--	--	70.11	8.88	--	--	74.87	9.58	66.52	11.13	58.08	14.01
WTBD JKT	65.65	12.46	65.68	12.51	58.99	12.46	69.23	10.58	58.61	13.12	68.75	10.98	62.59	12.90	59.44	12.62
Army-Wide PRS																
Effort	4.82	1.24	4.83	1.25	4.41	1.02	5.16	1.35	--	--	4.74	1.15	4.80	1.17	4.92	1.63
Physical Fitness & Bearing	4.76	1.22	4.79	1.19	4.44	1.17	5.03	1.35	--	--	4.68	1.18	4.71	1.11	4.72	1.76
Personal Discipline	4.90	1.27	4.96	1.21	4.42	1.14	5.26	1.49	--	--	4.81	1.21	4.80	1.18	4.75	1.74
Commitment & Adjustment	4.96	1.22	5.06	1.17	4.40	1.19	5.33	1.32	--	--	4.77	1.17	4.96	1.14	5.10	1.71
Support for Peers	4.95	1.18	4.99	1.16	4.51	1.14	5.31	1.38	--	--	4.82	1.11	4.94	1.09	5.31	1.42
Peer Leadership	4.67	1.29	4.57	1.29	4.49	1.37	4.97	1.44	--	--	4.67	1.24	4.74	1.13	5.17	1.37
Common Warrior Tasks KS	4.82	1.20	4.81	1.11	4.37	1.24	5.37	1.32	--	--	4.69	1.25	4.72	1.08	5.18	1.25
MOS Qualification KS	4.89	1.16	4.88	1.09	4.43	1.24	5.27	1.34	--	--	4.82	1.11	4.82	1.07	5.08	1.50
Overall Performance	3.50	0.84	3.45	.83	3.26	.79	3.58	.93	--	--	3.52	0.80	3.64	.78	3.41	1.15
MOS-Specific Performance Composite	4.55	0.95	4.70	0.89	--	--	4.63	1.01	--	--	4.35	0.84	4.92	0.90	4.84	1.75

Note. WTBD JKT test scores are percent correct, KS = Knowledge and Skills. A summary of the instrument sample sizes can be found in Table 2.5, while a summary of MOS frequencies can be found in Table 1.2. Results for 19K are not reported due to low sample size.

Table C.4. Army Life Questionnaire (ALQ) Intercorrelations for Full Schoolhouse Sample

Scale	1	2	3	4	5	6	7	8	9	10	11	12
1. Affective Commitment												
2. Normative Commitment	**.67**											
3. Career Intentions	**.57**	**.44**										
4. Reenlistment Intentions	**.55**	**.47**	**.86**									
5. Attrition Cognition	**-.62**	**-.74**	**-.47**	**-.50**								
6. Army Life Adjustment	**.45**	**.45**	**.37**	**.40**	**-.55**							
7. Army Civilian Comparison	**.43**	**.31**	**.33**	**.34**	**-.32**	**.23**						
8. MOS Fit	**.47**	**.40**	**.25**	**.27**	**-.42**	**.35**	**.24**					
9. Army Fit	**.83**	**.70**	**.55**	**.55**	**-.68**	**.61**	**.43**	**.48**				
10. Training Achievement	**.07**	**.02**	**.09**	**.07**	**-.05**	**.14**	**.02**	**.05**	**.08**			
11. Training Restarts	**-.07**	**-.08**	**-.02**	**-.03**	**.13**	**-.21**	.00	**-.08**	**-.10**	**-.11**		
12. Disciplinary Incidents	**-.09**	**-.10**	**-.05**	**-.07**	**.15**	**-.20**	-.01	**-.12**	**-.13**	**-.06**	**.17**	
13. Last APFT Score	**.05**	**.08**	**.02**	**.04**	**-.13**	**.25**	**-.02**	**.09**	**.11**	**.22**	**-.27**	**-.16**

Note. Significant correlation coefficients are bolded ($p < .05$). Sample sizes for each research criterion variable can be found in Table 2.5.

Table C.5. Correlations between the Army Life Questionnaire (ALQ) and Job Knowledge Tests (JKT) in Full Schoolhouse Sample

	Job Knowledge Tests						
ALQ Scales	Combined	11B	31B	68W	88M	91B	WTBD JKT
Affective Commitment	**.08**	**.10**	**.09**	**.07**	**.05**	**.10**	**.07**
Normative Commitment	**.17**	**.20**	**.14**	**.16**	**.13**	**.20**	**.17**
Career Intentions	**.02**	**.04**	.04	.01	-.03	.04	.00
Reenlistment Intentions	**.06**	**.07**	**.05**	**.05**	.03	**.10**	**.04**
Attrition Cognition	**-.15**	**-.16**	**-.15**	**-.16**	**-.11**	**-.15**	**-.16**
Army Life Adjustment	**.12**	**.10**	**.17**	**.13**	**.11**	**.10**	**.14**
Army Civilian Comparison	**.02**	**.04**	-.03	**.06**	**-.07**	.03	**-.02**
MOS Fit	**.11**	**.10**	**.05**	**.17**	.05	**.25**	**.13**
Army Fit	**.12**	**.15**	**.12**	**.13**	**.07**	**.13**	**.11**
Training Achievement	**-.11**	**-.16**	-.02	-.02	**-.15**	**-.12**	**-.08**
Training Restarts	**-.09**	**-.05**	**-.11**	**-.09**	**-.16**	**-.10**	**-.13**
Disciplinary Incidents	**-.04**	**-.04**	--	--	--	--	**-.07**
Last APFT Score	.01	**.03**	.03	.03	-.04	**-.13**	**.08**

Note. WTBD = Warrior Tasks and Battle Drills. Combined = MOS-specific JKT scores combined into one variable. Significant correlation coefficients are bolded ($p < .05$). Sample sizes for each research criterion variable can be found in Table 2.5.

Table C.6. Correlations between Army Life Questionnaire (ALQ) and Performance Rating Scales (PRS) in Full Schoolhouse Sample

	Army Life Questionnaire (ALQ)												
Performance Rating Scales	1	2	3	4	5	6	7	8	9	10	11	12	13
Army-Wide Performance Rating Scales													
Effort	**.11**	**.09**	**.04**	**.05**	**-.12**	**.13**	.02	**.09**	**.12**	**.09**	**-.11**	**-.15**	**.16**
Physical Fitness and Bearing	**.09**	**.08**	**.04**	**.05**	**-.13**	**.18**	**.03**	**.12**	**.12**	**.11**	**-.16**	**-.13**	**.28**
Personal Discipline	**.10**	**.08**	**.04**	**.05**	**-.11**	**.11**	.02	**.09**	**.11**	**.05**	**-.09**	**-.17**	**.12**
Commitment and Adjustment	**.11**	**.08**	**.06**	**.07**	**-.12**	**.14**	**.05**	**.09**	**.12**	**.07**	**-.12**	**-.14**	**.15**
Support for Peers	**.08**	**.07**	**.03**	**.05**	**-.09**	**.11**	**.03**	**.08**	**.09**	**.06**	**-.10**	**-.16**	**.13**
Peer Leadership	**.08**	**.07**	**.04**	**.05**	**-.11**	**.14**	.00	**.09**	**.10**	**.09**	**-.09**	**-.14**	**.18**
Common/Warrior Tasks KS	**.08**	**.06**	**.04**	**.05**	**-.09**	**.12**	**.03**	**.09**	**.09**	**.04**	**-.11**	**-.14**	**.14**
MOS Qualification KS	**.08**	**.06**	**.04**	**.05**	**-.10**	**.13**	**.03**	**.10**	**.08**	**.04**	**-.12**	**-.13**	**.15**
Overall Performance	**.08**	**.09**	**.03**	**.04**	**-.13**	**.18**	.00	**.11**	**.11**	**.13**	**-.14**	**-.16**	**.24**
MOS-Specific Performance Ratings Composite													
Combined MOS-Specific PRS	**.08**	**.04**	**.06**	**.06**	**-.07**	**.11**	.01	.02	**.08**	**.11**	**-.11**	**-.18**	**.09**
11B	**.11**	**.09**	**.08**	**.07**	**-.12**	**.15**	.00	**.14**	**.12**	**.12**	**-.14**	**-.15**	**.20**
31B	**.12**	**.12**	.01	.02	**-.19**	**.18**	-.02	.09	**.16**	**.13**	**-.17**	--	**.25**
68W	.01	.00	.03	.03	-.04	**.06**	**-.05**	.04	.01	**.05**	**-.08**	--	.04
88M	.02	-.01	.02	.01	-.03	.09	.04	.04	.03	**.14**	-.04	--	.12
91B	.12	.02	-.09	.04	-.04	.14	.05	.05	.09	-.07	.05	--	-.12

Note. KS = Knowledge and Skills. Significant correlation coefficients are bolded ($p < .05$). Sample sizes for each research criterion variable can be found in Table 2.5. 1=Affective Commitment; 2=Normative Commitment; 3=Career Intentions; 4=Reenlistment Intentions; 5=Attrition Cognition; 6=Army Life Adjustment; 7=Army Civilian Comparison; 8= MOS Fit; 9=Army Fit; 10=Training Achievement; 11=Training Restart; 12=Disciplinary Incidents; 13=Last APFT Score.

Table C.7 Correlations between Job Knowledge Tests (JKTs) and Performance Rating Scales (PRS) in Full Schoolhouse Sample

	MOS-Specific Job Knowledge Test (JKT)						
Performance Rating Scales	Total	11B	31B	68W	88M	91B	WTBD JKT
Army-Wide Performance Rating Scales							
Effort	**.06**	**.07**	**.16**	-.01	.09	.13	**.09**
Physical Fitness and Bearing	.03	.03	**.16**	-.01	.06	-.01	**.08**
Personal Discipline	**.05**	**.07**	**.12**	.00	.08	.03	**.10**
Commitment and Adjustment	**.03**	.03	**.12**	-.01	.03	.01	**.07**
Support for Peers	.02	**.04**	**.08**	**-.04**	.09	-.05	**.06**
Peer Leadership	.02	.03	**.10**	-.04	.06	-.07	**.07**
Common/Warrior Tasks Knowledge and Skills	.01	.04	.08	-.07	.03	-.11	**.06**
MOS Qualification Knowledge and Skills	.02	.04	**.08**	**-.06**	.01	.14	**.08**
Overall Performance	**.06**	**.08**	**.13**	.00	.02	.15	**.10**
MOS-Specific Performance Ratings Composite							
Combined MOS-Specific PRS	.00	-.03	**.14**	-.01	-.08	.00	**.02**
11B	.01	.01	--	--	--	--	.05
31B	**.14**	--	**.14**	--	--	--	**.10**
68W	-.01	--	--	-.01	--	--	.04
88M	-.08	--	--	--	-.08	--	-.01
91B	.00	--	--	--	--	.00	.15

Note. Significant correlation coefficients are bolded ($p < .05$). Sample sizes for each research criterion variable can be found in Table 2.5.

Table C.8. Descriptive Statistics for Administrative Criteria Based on the Validation Sample by MOS

	11B/11C/11X/18X			*19K*			*25U*			*31B*		
Administrative Criterion	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$
Attrtion[a]												
3-Month Cumulative	2,168	211	9.7	121	8	6.6	51	7	13.7	124	7	5.6
6-Month Cumulative	1,046	123	11.8	108	14	13.0	33	6	18.2	75	13	17.3
Initial Military Training (IMT) Criteria	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$
Restarted at Least Once During IMT	1,764	240	13.6	119	19	16.0	48	3	6.3	274	33	12.0
Restarted at Least Once During IMT for Pejorative Reasons	1,761	237	13.5	117	17	14.5	48	3	6.3	272	31	11.4
Restarted at Least Once During IMT for Academic Reasons	1,624	100	6.2	105	5	4.8	47	2	4.3	249	8	3.2
AIT School Grades	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*
Overall Average (Unstandardized)	--	--	--	--	--	--	61	94.47	3.37	--	--	--
Overall Average (Standardized within MOS)	--	--	--	--	--	--	61	0.12	0.96	--	--	--

	42A			*68W*			*88M*			*91B*		
Administrative Criterion	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$
Attrtion[a]												
3-Month Cumulative	70	5	0.9	475	26	1.5	389	28	7.2	235	20	8.5
6-Month Cumulative	39	3	0.6	253	20	1.2	133	15	11.3	112	12	10.7
Initial Military Training (IMT) Criteria	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$
Restarted at Least Once During IMT	121	6	5.0	175	26	14.9	317	72	22.7	137	10	7.3
Restarted at Least Once During IMT for Pejorative Reasons	121	6	5.0	168	19	11.3	260	14	5.4	134	7	5.2
Restarted at Least Once During IMT for Academic Reasons	119	4	3.4	175	26	14.9	316	71	22.5	136	9	6.6
AIT School Grades	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*
Overall Average (Unstandardized)	--	--	--	85	85.65	9.13	--	--	--	--	--	--
Overall Average (Standardized within MOS)	--	--	--	85	0.02	1.03	--	--	--	--	--	--

Note. Results are limited to non-prior service, Education Tier 1, AFQT Category IV or higher.

[a] Attrition results reflect Regular Army Soldiers only.

[b] N = number of Soldiers with 3-month attrition data at the time data were extracted. N_{Attrit} = number of Soldiers who attrited through 3 months of service. $\%_{Attrit}$ = percentage of Soldiers who attrited through 3 months of service [(N_{Attrit} /N) x 100].

[c] N = number of Soldiers with ATRRS IMT data at the time data were extracted. $N_{Restart}$ = number of Soldiers who restarted at least once during IMT. $\%_{Restart}$ = percentage of Soldiers who restarted at least once during IMT [($N_{Restart}$ /N) x 100].

[d] N = number of Soldiers with RITMS AIT school grade data. Standardized school grades were not computed for MOS with insufficient sample size (n < 15).

APPENDIX D: SUPPLEMENTAL VALIDITY TABLES

Table D.1. Incremental Validity Estimates for the TAPAS Scales over the AFQT for Predicting Performance- and Retention-Related Criteria

Criterion	*n*	AFQT Only *R* (r_{pb})	AFQT + TAPAS *R* (r_{pb})	Δ*R* (Δr_{pb})
Can-Do Performance				
WTBD JKT	1,992	**.493**	**.499**	.005
MOS-Specific JKT	1,661	**.378**	**.392**	.014
MOS Proficiency (PRS)	626	.030	.189	.159
MOS-Specific PRS	478	.044	.160	.116
IMT Exam Grade	2,920	**.299**	**.313**	**.014**
Graduated IMT without Restart	3,058	**(.040)**	**(.098)**	**(.059)**
Training Achievement (ALQ)	2,036	**.128**	**.225**	**.097**
Training Restarts (ALQ)	2,040	**.084**	**.243**	**.159**
Common/Warrior Tasks KS (PRS)	631	.037	.175	.138
Will-Do Performance				
Exhibiting Effort (PRS)	649	.046	.195	.149
Support for Peers (PRS)	645	.027	.196	.169
Peer Leadership (PRS)	621	.028	.168	.141
Exhibiting Fitness and Bearing (PRS)	650	.037	**.205**	**.168**
Personal Discipline (PRS)	654	**.089**	**.223**	**.133**
Last APFT Score (ALQ)	2,018	**.108**	**.320**	**.211**
Disciplinary Action (ALQ)	943	.044	.149	.104
Commitment and Adjustment (PRS)	651	.010	.168	.157
Retention				
Adjustment to Army Life (ALQ)	2,040	**.086**	**.262**	**.175**
Affective Commitment (ALQ)	2,040	**.090**	**.235**	**.145**
Normative Commitment (ALQ)	2,040	**.083**	**.177**	**.094**
Career Intentions (ALQ)	2,040	**.140**	**.199**	**.059**
Attrition Cognitions (ALQ)	2,040	.024	**.181**	**.157**
Reenlistment Intentions (ALQ)	2,040	**.085**	**.172**	**.088**
Army Fit (ALQ)	2,040	.042	**.226**	**.184**
MOS Fit (ALQ)	2,040	**.049**	**.149**	**.099**
Army Civilian Comparison (ALQ)	2,040	**.189**	**.206**	.017
3-Month Attrition[a]	8,242	**(.058)**	**(.105)**	**(.047)**
6-Month Attrition[a]	3,439	**(.062)**	**(.123)**	**(.061)**
Overall Performance (PRS)	649	**.084**	**.212**	.128

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. PRS = Performance Rating Scales. AFQT Only = Correlation between the AFQT and the criterion of interest. AFQT + TAPAS = Multiple correlation (*R*) between the AFQT and the selected predictor measure with the criterion of interest. Δ*R* = Increment in *R* over the AFQT from adding the selected predictor measure to the regression model ([AFQT + TAPAS] – AFQT Only). Estimates in parentheses are *point-biserial correlations* (r_{pb}) that reflect the observed point-biserial correlation between Soldiers' predicted probability of attriting and their actual attrition behavior. Large, positive r_{pb} values mean that the TOPS composite or scale performed well in predicting actual attrition. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a]Attrition results include Regular Army Soldiers only.

Table D.2. Bivariate and Semi-Partial Correlations between the TAPAS Scales and Can-Do Performance-Related Criteria

	Criteria								
	WTBD JKT	MOS-Specific JKT	MOS Proficiency (PRS)	MOS-Specific PRS	IMT Exam Grade	Graduated IMT without Restart	Training Achievement (ALQ)	Training Restarts (ALQ)	Common/ Warrior Tasks KS (PRS)
TAPAS Dimensions	*N* = 2,100	*N* = 1,751	*N* = 666	*N* = 515	*N* = 3,098	*N* = 5,644	*N* = 2,147	*N* = 2,151	*N* = 677
Achievement	**.05** (.00)	.01 (-.03)	.03 (.02)	-.03 (-.02)	.03 (.00)	.01 (.00)	**.08 (.09)**	**-.09 (-.08)**	.02 (.02)
Adjustment[a]	**.08** (.03)	.03 (-.01)	-.05 (-.05)	-.03 (-.03)	.00 **(-.04)**	-.02 (-.02)	.02 (.04)	**-.05** (-.04)	-.01 (-.02)
Attention Seeking	.03 (-.02)	.03 (-.01)	.07 (.07)	.03 (.03)	-.03 **(-.06)**	**.04 (.03)**	**.05 (.06)**	-.03 (-.02)	.07 (.07)
Cooperation	-.02 (-.02)	.01 (.01)	-.03 (-.03)	-.02 (-.02)	-.03 (-.03)	-.01 (-.01)	**-.06 (-.06)**	**.04 (.04)**	-.01 (-.01)
Dominance	.03 (-.01)	-.03 **(-.07)**	.07 (.07)	.00 (.01)	.01 (-.02)	**.04 (.04)**	**.10 (.11)**	**-.09 (-.08)**	.06 (.05)
Even Tempered	.02 (-.02)	.03 (.00)	-.02 (-.02)	-.04 (-.03)	.02 (-.01)	**-.04 (-.04)**	-.04 (-.03)	.01 (.02)	.04 (.04)
Generosity	**-.05** (-.01)	**-.05** (-.03)	-.01 (.00)	-.05 (-.05)	-.03 (-.01)	.00 (.01)	-.01 (-.02)	**.04** (.04)	-.02 (-.01)
Intellectual Efficiency	**.23** (.02)	**.14** (-.02)	.02 (.01)	-.08 (-.06)	**.12** (-.01)	.00 (-.02)	-.04 (.02)	**-.10 (-.07)**	-.01 (-.03)
Non-delinquency	-.03 (-.03)	-.01 (-.01)	-.02 (-.02)	-.03 (-.03)	.03 **(.04)**	-.02 (-.02)	-.03 (-.03)	**.06 (.06)**	.04 (.04)
Optimism	.00 (.00)	.01 (.01)	.04 (.04)	-.01 (.00)	-.02 (-.02)	.01 (.01)	.01 (.01)	-.01 (-.01)	.03 (.03)
Order	**-.08** (.00)	**-.11** (-.04)	**.11 (.12)**	.07 (.06)	-.01 **(.04)**	-.01 (.00)	**.08 (.06)**	.00 (-.01)	**.09 (.10)**
Physical Conditioning	.03 (.00)	-.01 (-.03)	**.09 (.09)**	.06 (.06)	.00 (-.02)	**.08 (.07)**	**.11 (.12)**	**-.18 (-.18)**	.05 (.05)
Self Control[a]	-.01 (-.01)	-.06 **(-.05)**	.02 (.02)	.03 (.03)	.01 (.01)	**-.04 (-.04)**	.02 (.01)	.01 (.01)	.04 (.04)
Sociability	**-.07** (-.03)	**-.07** (-.04)	.00 (.00)	-.03 (-.03)	**-.06 (-.04)**	.02 (.02)	.02 (.01)	.01 (.01)	-.01 (-.01)
Tolerance	-.04 (-.03)	-.02 (-.01)	-.01 (-.01)	-.05 (-.05)	**-.04** (-.03)	-.01 (-.01)	.00 (.00)	**.08 (.08)**	.00 (.00)
TAPAS Composites									
Can-Do Composite	**.09** (-.02)	**.06** (-.02)	.02 (.01)	-.06 (-.05)	**.07** (.00)	-.02 **(-.03)**	.00 (.03)	**-.04** (-.03)	.05 (.04)
Will-Do Composite	.00 (-.02)	-.01 (-.03)	.00 (.00)	-.03 (-.02)	**.05** (.03)	-.01 (-.01)	.04 (.04)	**-.07 (-.06)**	.03 (.03)

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in parentheses are semi-partial correlations between the TAPAS scales and the criterion of interest, controlling for AFQT. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 478 – 17,993.

Table D.3. Bivariate and Semi-Partial Correlations between the TAPAS Scales and Will-Do Performance-Related Criteria

	Criteria																	
	Exhibiting Effort (PRS)		Support for Peers (PRS)		Peer Leadership (PRS)		Exhibiting Fitness & Bearing (PRS)		Personal Discipline (PRS)		Last APFT Score (ALQ)		Disciplinary Incidents (ALQ)		Commitment & Adjustment (PRS)	
TAPAS Dimensions	*N* = 696		*N* = 694		*N* = 668		*N* = 699		*N* = 703		*N* = 2,128		*N* = 996		*N* = 700	
Achievement	.05	(.05)	.00	(.00)	.05	(.05)	.04	(.04)	.06	(.05)	**.08**	**(.06)**	**-.10**	**(-.09)**	.04	(.04)
Adjustment[a]	-.03	(-.03)	**-.09**	**(-.09)**	-.05	(-.05)	-.05	(-.05)	-.04	(-.05)	.01	(.00)	-.02	(-.02)	-.03	(-.03)
Attention Seeking	**.08**	**(.08)**	**.10**	**(.10)**	.06	(.06)	**.09**	**(.08)**	**.08**	(.07)	.03	(.02)	.03	(.04)	**.08**	**(.07)**
Cooperation	.02	(.02)	-.05	(-.05)	-.04	(-.04)	-.01	(-.01)	-.07	(-.07)	-.02	(-.02)	.01	(.01)	-.05	(-.05)
Dominance	.06	(.06)	.03	(.03)	.02	(.02)	.06	(.06)	.02	(.01)	**.13**	**(.12)**	-.04	(-.03)	.06	(.06)
Even Tempered	.05	(.05)	.01	(.01)	.00	(.00)	.04	(.04)	.05	(.04)	**-.08**	**(-.09)**	.03	(.03)	.00	(-.01)
Generosity	-.05	(-.04)	-.04	(-.04)	-.02	(-.01)	-.03	(-.02)	-.03	(-.02)	.01	(.02)	-.06	(-.06)	-.04	(-.04)
Intellectual Efficiency	-.01	(-.04)	-.03	(-.04)	.01	(.00)	-.04	(-.06)	.01	(-.03)	.04	(-.01)	-.02	(.00)	-.02	(-.02)
Non-delinquency	.01	(.01)	.02	(.02)	.03	(.03)	-.03	(-.03)	-.01	(-.01)	**-.08**	**(-.08)**	-.01	(-.01)	.00	(.00)
Optimism	.03	(.03)	.02	(.02)	.05	(.05)	.04	(.04)	.05	(.04)	.02	(.02)	-.01	(-.01)	.05	(.05)
Order	.06	(.07)	.06	(.07)	**.10**	**(.11)**	.06	(.07)	**.09**	**(.11)**	.03	**(.05)**	-.01	(-.02)	.06	(.06)
Physical Conditioning	**.11**	**(.11)**	.03	(.02)	.05	(.04)	**.11**	**(.11)**	.02	(.02)	**.27**	**(.27)**	-.06	(-.05)	.07	(.07)
Self Control[a]	.01	(.02)	-.01	(-.01)	-.02	(-.02)	-.01	(-.01)	.03	(.03)	-.02	(-.02)	.01	(.01)	.00	(.00)
Sociability	.00	(.00)	-.03	(-.03)	.00	(.01)	.01	(.01)	-.05	(-.05)	.02	(.03)	.02	(.01)	-.03	(-.03)
Tolerance	.01	(.01)	.01	(.01)	-.01	(-.01)	.02	(.02)	.01	(.02)	.02	(.02)	-.02	(-.02)	-.02	(-.02)
TAPAS Composites																
Can-Do Composite	.04	(.03)	.00	(.00)	.05	**(.05)**	.02	(.01)	.05	(.03)	-.01	(-.03)	-.04	(-.03)	.02	(.02)
Will-Do Composite	.05	(.05)	-.03	(-.03)	.03	(.03)	.03	(.03)	.01	(.01)	**.06**	**(.06)**	**-.07**	**(-.06)**	.01	(.01)

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in parentheses are semi-partial correlations between the TAPAS scales and the criterion of interest, controlling for AFQT. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 621 – 2,018.

Table D.4. Bivariate and Semi-Partial Correlations between the TAPAS Scales and Retention-Related Criteria

	Criteria																			
	Adjustment to Army Life (ALQ)		Affective Commitment (ALQ)		Normative Commitment (ALQ)		Career Intentions (ALQ)		Attrition Cognitions (ALQ)		Reenlistment Intentions (ALQ)		Army Fit (ALQ)		MOS Fit (ALQ)		Army Civilian Comparison (ALQ)		3-Month Attrition[b]		6-Month Attrition[b]	
TAPAS Dimensions	*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 2,151		*N* = 8,638		*N* = 3,810	
Achievement	**.13**	**(.12)**	**.12**	**(.13)**	**.12**	**(.11)**	**.09**	**(.10)**	**-.14**	**(-.14)**	**.10**	**(.11)**	**.15**	**(.15)**	**.08**	(.08)	.00	(.02)	.00	(.01)	.00	(.00)
Adjustment[a]	**.09**	**(.08)**	**-.05**	(-.04)	-.01	(-.02)	-.01	(.01)	.00	(.00)	-.01	(.00)	.00	(.01)	.04	**(.03)**	-.02	(.00)	-.02	(-.01)	-.01	(.00)
Attention Seeking	**.05**	**(.04)**	**.07**	**(.08)**	.04	(.03)	.01	(.02)	**-.05**	**(-.05)**	-.01	(.00)	**.06**	**(.07)**	.03	(.02)	-.02	(.00)	-.01	(-.01)	**-.04**	(-.03)
Cooperation	-.03	(-.03)	.00	(.00)	-.01	(-.01)	.01	(.01)	.00	(.00)	.01	(.01)	-.01	(-.01)	-.02	(-.02)	.02	(.02)	-.01	(-.01)	.00	(.00)
Dominance	**.15**	**(.14)**	**.12**	**(.13)**	**.08**	**(.07)**	**.06**	**(.08)**	**-.09**	**(-.09)**	**.07**	**(.07)**	**.13**	**(.14)**	**.07**	**(.07)**	-.04	(-.03)	.00	(.01)	-.02	**(-.01)**
Even Tempered	.03	(.03)	.00	(.01)	.04	(.03)	.04	**(.05)**	**-.04**	(-.04)	**.05**	**(.06)**	.02	(.02)	.00	(.00)	.00	(.02)	-.01	(-.01)	.00	(.00)
Generosity	.00	(.00)	**.08**	**(.07)**	**.05**	**(.06)**	**.07**	**(.06)**	-.03	(-.04)	**.06**	**(.06)**	**.09**	**(.08)**	.00	(.01)	.01	(-.01)	**.04**	(.03)	.02	(.02)
Intellectual	**.12**	**(.10)**	-.02	(.02)	**.06**	(.03)	-.01	**(.05)**	**-.05**	(-.04)	.00	(.04)	.02	**(.05)**	.03	(.01)	**-.10**	(-.02)	-.01	**(.02)**	-.01	**(.02)**
Efficiency																						
Non-delinquency	.02	(.02)	**.05**	**(.05)**	.03	(.03)	.04	(.04)	-.03	(-.03)	.03	(.03)	.04	(.04)	.02	(.02)	.04	(.04)	.01	(.01)	.02	(.02)
Optimism	**.12**	**(.11)**	**.08**	**(.08)**	.02	(.01)	**.07**	**(.07)**	**-.06**	**(-.06)**	**.07**	**(.07)**	**.08**	**(.09)**	**.05**	**(.05)**	.02	(.02)	**-.03**	(-.03)	**-.04**	**(-.04)**
Order	.03	**(.04)**	.01	(-.01)	.00	(.01)	**.05**	(.03)	.00	(.00)	.04	(.03)	.03	(.02)	**-.05**	**(-.05)**	.02	(-.02)	**.02**	**(.01)**	.03	(.02)
Physical	**.18**	**(.17)**	**.06**	**(.06)**	**.05**	**(.05)**	.01	(.01)	**-.09**	**(-.09)**	.00	(.01)	**.08**	**(.08)**	**.09**	**(.08)**	.00	(.01)	**-.05**	(-.05)	**-.07**	**(-.07)**
Conditioning																						
Self Control[a]	.02	(.03)	**.05**	**(.04)**	.03	(.03)	**.05**	**(.05)**	-.03	(-.03)	**.05**	**(.05)**	**.05**	**(.05)**	.00	(.00)	**.05**	**(.05)**	.00	**(.00)**	.01	(.01)
Sociability	.02	(.03)	.03	(.02)	-.03	(-.02)	.02	(.01)	.01	(.00)	.02	(.01)	.01	(.01)	**.05**	**(.05)**	.01	(-.01)	.00	(-.01)	-.02	(-.02)
Tolerance	.03	(.03)	.04	(.04)	.02	(.03)	**.06**	**(.05)**	-.04	(-.04)	**.06**	**(.06)**	**.05**	**(.05)**	.02	(.02)	.02	(.02)	.00	(.00)	.02	(.02)
TAPAS Composites																						
Can-Do Composite	**.15**	**(.14)**	**.09**	**(.11)**	**.09**	(.08)	**.08**	**(.11)**	**-.11**	**(-.11)**	**.09**	**(.11)**	**.11**	**(.12)**	**.07**	**(.06)**	-.02	(.03)	-.01	(.00)	-.02	(.00)
Will-Do Composite	**.13**	**(.12)**	**.07**	**(.07)**	**.08**	(.08)	**.07**	**(.08)**	**-.10**	**(-.10)**	**.08**	**(.08)**	**.09**	**(.09)**	**.06**	**(.06)**	.03	(.04)	-.01	(-.01)	-.01	(-.01)

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in parentheses are semi-partial correlations between the TAPAS scales and the criterion of interest, controlling for AFQT. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 2,040 – 16,475

[b] Attrition results include Regular Army Soldiers only.